NUCLEIC ACID SEQUENCING USING MICROSPHERE ARRAYS

This application is a continuation-in-part application of U.S.S.N.s 60/130,089 filed April 20, 1999; 60/135,051, filed May 20, 1999; 60/135,053, filed May 20, 1999; 60/135,123, filed May 20, 1999; and 60/160,027, filed October 22, 1999. It also claims priority to 60/161,148; filed October 22, 1999; 09/324,633, filed October 22, 1999; and 60/160,917, filed October 22, 1999.

FIELD OF THE INVENTION

The invention relates to DNA sequencing by synthesis techniques, including those utilizing the detection of pyrophosphate (PPi) generated during the DNA synthesis reaction (pyrosequencing). The methods and compositions utilize biosensor arrays, particularly microsphere arrays.

BACKGROUND OF THE INVENTION

DNA sequencing is a crucial technology in biology today, as the rapid sequencing of genomes, including the human genome, is both a significant goal and a significant hurdle. Thus there is a significant need for robust, high-throughput methods. Traditionally, the most common method of DNA sequencing has been based on polyacrylamide gel fractionation to resolve a population of chain-terminated fragments (Sanger et al., Proc. Natl. Acad. Sci. USA 74:5463 (1977); Maxam & Gilbert). The population of fragments, terminated at each position in the DNA sequence, can be generated in a number of ways. Typically, DNA polymerase is used to incorporate dideoxynucleotides that serve as chain terminators.

Several alternative methods have been developed to increase the speed and ease of DNA sequencing. For example, sequencing by hybridization has been described (Drmanac et al., Genomics 4:114 (1989); Koster et al., Nature Biotechnology 14:1123 (1996); U.S. Patent Nos. 5,525,464; 5,202,231 and 5,695,940, among others). Similarly, sequencing by synthesis is an alternative to gel-based sequencing. These methods add and read only one base (or at most a few bases, typically of the same type) prior to polymerization of the next base. This can be referred to as

"time resolved" sequencing, to contrast from "gel-resolved" sequencing. Sequencing by synthesis has been described in U. S. Patent No 4,971,903 and Hyman, Anal. Biochem. 174:423 (1988); Rosenthal, International Patent Application Publication 761107 (1989); Metzker et al., Nucl. Acids Res. 22:4259 (1994); Jones, Biotechniques 22:938 (1997); Ronaghi et al., Anal. Biochem. 242:84 (1996), Nyren et al., Anal. Biochem. 151:504 (1985). Detection of ATP sulfurylase activity is described in Karamohamed and Nyren, Anal. Biochem. 271:81 (1999). Sequencing using reversible chain terminating nucleotides is described in U.S. Patent Nos. 5,902,723 and 5,547,839, and Canard and Arzumanov, Gene 11:1 (1994), and Dyatkina and Arzumanov, Nucleic Acids Symp Ser 18:117 (1987). Reversible chain termination with DNA ligase is described in U.S. Patent 5,403,708. Time resolved sequencing is described in Johnson et al., Anal. Biochem. 136:192 (1984). Single molecule analysis is described in U.S. Patent No. 5,795,782 and Elgen and Rigler, Proc. Natl Acad Sci USA 91(13):5740 (1994), all of which are hereby expressly incorporated by reference in their entirety.

One promising sequencing by synthesis method is based on the detection of the pyrophosphate (PPi) released during the DNA polymerase reaction. As nucleotriphosphates are added to a growing nucleic acid chain, they release PPi. This release can be quantitatively measured by the conversion of PPi to ATP by the enzyme sulfurylase, and the subsequent production of visible light by firefly luciferase.

Several assay systems have been described that capitalize on this mechanism. See for example WO93/23564, WO 98/28440 and WO98/13523, all of which are expressly incorporated by reference. A preferred method is described in Ronaghi et al., Science 281:363 (1998). In this method, the four deoxynucleotides (dATP, dGTP, dCTP and dTTP; collectively dNTPs) are added stepwise to a partial duplex comprising a sequencing primer hybridized to a single stranded DNA template and incubated with DNA polymerase, ATP sulfurylase, luciferase, and optionally a nucleotide-degrading enzyme such as apyrase. A dNTP is only incorporated into the growing DNA strand if it is complementary to the base in the template strand. The synthesis of DNA is accompanied by the release of PPi equal in molarity to the incorporated dNTP. The PPi is converted to ATP and the light generated by the luciferase is directly proportional to the amount of ATP. In some cases the unincorporated dNTPs and the produced ATP are degraded between each cycle by the nucleotide degrading enzyme.

In some cases the DNA template is associated with a solid support. To this end, there are a wide variety of known methods of attaching DNAs to solid supports. Recent work has focused on the attachment of binding ligands, including nucleic acid probes, to microspheres that are randomly distributed on a surface, including a fiber optic bundle, to form high density arrays. See for example PCTs US98/21193, PCT US99/14387 and PCT US98/05025; WO98/50782; and U.S.S.N.s 09/287,573, 09/151,877, 09/256,943, 09/316,154, 60/119,323, 09/315,584; all of which are expressly incorporated by reference.

Accordingly, it is an object of the invention to provide compositions and methods of sequencing nucleic acids using arrays.

SUMMARY OF INVENTION

In accordance with the above identified objects, the present invention provides methods of sequencing a plurality of target nucleic acids. The methods comprise providing a plurality of hybridization complexes each comprising a target sequence and a sequencing primer that hybridizes to the first domain of the target sequence, the hybridization complexes are attached to a surface of a substrate. The methods comprise extending each of the primers by the addition of a first nucleotide to the first detection position using an enzyme to form an extended primer. The methods comprise detecting the release of pyrophosphate (PPi) to determine the type of the first nucleotide added onto the primers. In one aspect the hybridization complexes are attached to microspheres distributed on the surface. In an additional aspect the sequencing primers are attached to the surface. The hybridization complexes comprise the target sequence, the sequencing primer and a capture probe covalently attached to the surface. The hybridization complexes also comprise an adapter probe.

In an additional aspect, the method comprises extending the extended primer by the addition of a second nucleotide to the second detection position using an enzyme and detecting the release of pyrophosphate to determine the type of second nucleotide added onto the primers. In an additional aspect, the pyrophosphate is detected by contacting the pyrophosphate with a second enzyme that converts pyrophosphate into ATP, and detecting the ATP using a third enzyme. In one aspect, the second enzyme is sulfurylase and/or the third enzyme is luciferase.

In an additional aspect, the invention provides methods of sequencing a target nucleic acid comprising a first domain and an adjacent second domain, the second domain comprising a plurality of target positions. The method comprises providing a hybridization complex comprising the target sequence and a capture probe covalently attached to microspheres on a surface of a substrate and determining the identity of a plurality of bases at the target positions. The hybridization complex comprises the capture probe, an adapter probe, and the target sequence. In one aspect the sequencing primer is the capture probe.

In an additional aspect of the invention, the determining comprises providing a sequencing primer hybridized to the second domain, extending the primer by the addition of first nucleotide to the first detection position using a first enzyme to form an extended primer, detecting the release of pyrophosphate to determine the type of the first nucleotide added onto the primer, extending the primer by the addition of a second nucleotide to the second detection position using the enzyme, and

detecting the release of pyrophosphate to determine the type of the second nucleotide added onto the primer. In an additional aspect pyrophosphate is detected by contacting the pyrophosphate with the second enzyme that converts pyrophosphate into ATP, and detecting the ATP using a third enzyme. In one aspect the second enzyme is sulfurylase and/or the third enzyme is luciferase.

In an additional aspect of the method for sequencing, the determining comprises providing a sequencing primer hybridized to the second domain, extending the primer by the addition of a first protected nucleotide using a first enzyme to form an extended primer, determining the identification of the first protected nucleotide, removing the protection group, adding a second protected nucleotide using the enzyme, and determining the identification of the second protected nucleotide.

In an additional aspect the invention provides a kit for nucleic acid sequencing comprising a composition comprising a substrate with a surface comprising discrete sites and a population of microspheres distributed on the sites, wherein the microspheres comprise capture probes. The kit also comprises an extension enzyme and dNTPs. The kit also comprises a second enzyme for the conversion of pyrophosphate to ATP and a third enzyme for the detection of ATP. In one aspect the dNTPs are labeled. In addition each dNTP comprises a different label.

BRIEF DESCRIPTION OF THE FIGURES

Figures 1A, 1B, 1C and 1D depict several configurations for attachment of the target sequences to the arrays of the invention. Bead arrays are depicted, although as outlined herein, any number of additional arrays may be used. Figure 1A depicts a substrate **5** with a capture probe **20** attached via an optional attachment linker **15** to an associated microsphere **10**. Target sequence **25** comprises target positions **30, 31, 32,** and **33** with a sequencing primer **40** hybridized adjacently to these positions. There may be any number of sets of target positions ($n \geq 1$). Figure 1B depicts the use of the capture probe **20** as the sequencing primer. Figure 1C depicts the use of a capture extender probe (sometimes referred to herein as an "adapter probe") **50** that has a first domain that hybridizes to the capture probe **20** and a second portion that hybridizes to the target sequence **25**. Figure 1D shows the direct attachment of the target sequence **25** to the bead **10**.

DETAILED DESCRIPTION

The present invention is directed to the sequencing of nucleic acids, particularly DNA, by synthesizing nucleic acids using the target sequence (i.e. the nucleic acid for which the sequence is determined) as a template. These methods can be generally described as follows. A target sequence is attached to a solid support, either directly or indirectly, as outlined below. The target sequence comprises a first

domain and an adjacent second domain comprising target positions for which sequence information is desired. A sequencing primer is hybridized to the first domain of the target sequence, and an extension enzyme is added, such as a polymerase or a ligase, as outlined below. After the addition of each base, the identity of each newly added base is determined prior to adding the next base. This can be done in a variety of ways, including controlling the reaction rate and using a fast detector, such that the newly added bases are identified in real time. Alternatively, the addition of nucleotides is controlled by reversible chain termination, for example through the use of photocleavable blocking groups. Alternatively, the addition of nucleotides is controlled, so that the reaction is limited to one or a few bases at a time. The reaction is restarted after each cycle of addition and reading. Alternatively, the addition of nucleotides is accomplished by carrying out a ligation reaction with oligonucleotides comprising chain terminating oligonucleotides. Preferred methods of sequencing-by-synthesis include, but are not limited to, pyrosequencing, reversible-chain termination sequencing, time-resolved sequencing, ligation sequencing, and single-molecule analysis, all of which are described below.

The advantages of these "sequencing-by-synthesis" reactions can be augmented through the use of array techniques that allow very high density arrays to be made rapidly and inexpensively, thus allowing rapid and inexpensive nucleic acid sequencing. By "array techniques" is meant techniques that allow for analysis of a plurality of nucleic acids in an array format. The maximum number of nucleic acids is limited only by the number of discrete loci on a particular array platform. As is more fully outlined below, a number of different array formats can be used.

The methods of the invention find particular use in sequencing a target nucleic acid sequence, i.e. identifying the sequence of a target base or target bases in a target nucleic acid, which can ultimately be used to determine the sequence of long nucleic acids.

Accordingly, the present invention provides methods of sequencing target nucleic acids in sample solutions. As will be appreciated by those in the art, the sample solution may comprise any of a number of things, including, but not limited to, bodily fluids (including, but not limited to, blood, urine, serum, lymph, saliva, anal and vaginal secretions, perspiration and semen, of virtually any organism, with mammalian samples being preferred and human samples being particularly preferred); environmental samples (including, but not limited to, air, agricultural, water and soil samples); biological warfare agent samples; research samples (i.e. in the case of nucleic acids, the sample may be the products of an amplification reaction, including both target and signal amplification as is generally described in "Detection of Nucleic Acid Amplification Reactions Using Bead Arrays", filed October 22, 1999, U.S.S.N. 60/161,048 hereby incorporated by reference, such as PCR amplification reaction); purified samples, such as purified genomic DNA, RNA, proteins, etc.; raw samples (bacteria, virus, genomic DNA, etc.; as will be appreciated by those in the art, virtually any experimental

manipulation may have been done on the sample.

If required, the target sequence is prepared using known techniques. For example, the sample may be treated to lyse the cells, using known lysis buffers, electroporation, etc., with purification and/or amplification as needed, as will be appreciated by those in the art. Suitable amplification techniques are outlined in "Detection of Nucleic Acid Amplification Reactions Using Bead Arrays", filed October 22, 1999, U.S.S.N. 60/161,048, hereby expressly incorporated by reference. However, in some embodiments, no purification or amplification is necessary. As will be appreciated by those in the art, the target sequences may comprise both single-stranded and double-stranded portions, although the portion to which the sequencing primer hybridizes must be single-stranded. This single-stranded portion may be generated either before or after array synthesis. Similarly, a preferred embodiment has a single-stranded extension area (i.e. the sequence that is generated by the enzyme and read), although in some instances, the enzyme that extends the primer, i.e. the DNA polymerase, will displace or degrade a second strand. In some cases, a primer need not be used; for example, as described in Ronaghi et al., supra, a T7 RNA polymerase promoter may be used to direct synthesis using T7 RNA polymerase.

The present invention provides compositions comprising arrays with attached nucleic acids and methods for identifying the sequence of nucleic acids. By "nucleic acid" or "oligonucleotide" or grammatical equivalents herein means at least two nucleotides covalently linked together. A nucleic acid of the present invention will generally contain phosphodiester bonds, although in some cases, as outlined below, nucleic acid analogs are included that may have alternate backbones, comprising, for example, phosphoramide (Beaucage et al., Tetrahedron 49(10):1925 (1993) and references therein; Letsinger, J. Org. Chem. 35:3800 (1970); Sprinzl et al., Eur. J. Biochem. 81:579 (1977); Letsinger et al., Nucl. Acids Res. 14:3487 (1986); Sawai et al, Chem. Lett. 805 (1984), Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); and Pauwels et al., Chemica Scripta 26:141 91986)), phosphorothioate (Mag et al., Nucleic Acids Res. 19:1437 (1991); and U.S. Patent No. 5,644,048), phosphorodithioate (Briu et al., J. Am. Chem. Soc. 111:2321 (1989), O-methylphophoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press), and peptide nucleic acid backbones and linkages (see Egholm, J. Am. Chem. Soc. 114:1895 (1992); Meier et al., Chem. Int. Ed. Engl. 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson et al., Nature 380:207 (1996), all of which are incorporated by reference). Other analog nucleic acids include those with positive backbones (Denpcy et al., Proc. Natl. Acad. Sci. USA 92:6097 (1995); non-ionic backbones (U.S. Patent Nos. 5,386,023, 5,637,684, 5,602,240, 5,216,141 and 4,469,863; Kiedrowshi et al., Angew. Chem. Intl. Ed. English 30:423 (1991); Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); Letsinger et al., Nucleoside & Nucleotide 13:1597 (1994); Chapters 2 and 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y.S. Sanghui and P. Dan Cook;

Mesmaeker et al., Bioorganic & Medicinal Chem. Lett. 4:395 (1994); Jeffs et al., J. Biomolecular NMR 34:17 (1994); Tetrahedron Lett. 37:743 (1996)) and non-ribose backbones, including those described in U.S. Patent Nos. 5,235,033 and 5,034,506, and Chapters 6 and 7, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y.S. Sanghui and P. Dan Cook. Nucleic acids containing one or more carbocyclic sugars are also included within the definition of nucleic acids (see Jenkins et al., Chem. Soc. Rev. (1995) pp169-176). Several nucleic acid analogs are described in Rawls, C & E News June 2, 1997 page 35. All of these references are hereby expressly incorporated by reference. These modifications of the ribose-phosphate backbone may be done to facilitate the addition of labels, or to increase the stability and half-life of such molecules in physiological environments.

As will be appreciated by those in the art, all of these nucleic acid analogs may find use in the present invention. In addition, mixtures of naturally occurring nucleic acids and analogs can be made. Alternatively, mixtures of different nucleic acid analogs, and mixtures of naturally occuring nucleic acids and analogs may be made.

Particularly preferred are peptide nucleic acids (PNA) which includes peptide nucleic acid analogs. These backbones are substantially non-ionic under neutral conditions, in contrast to the highly charged phosphodiester backbone of naturally occurring nucleic acids. This results in two advantages. First, the PNA backbone exhibits improved hybridization kinetics. PNAs have larger changes in the melting temperature (Tm) for mismatched versus perfectly matched basepairs. DNA and RNA typically exhibit a 2-4°C drop in Tm for an internal mismatch. With the non-ionic PNA backbone, the drop is closer to 7-9°C. This allows for better detection of mismatches. Similarly, due to their non-ionic nature, hybridization of the bases attached to these backbones is relatively insensitive to salt concentration.

The nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. The nucleic acid may be DNA, both genomic and cDNA, RNA or a hybrid, where the nucleic acid contains any combination of deoxyribo- and ribo-nucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xathanine hypoxathanine, isocytosine, isoguanine, etc. A preferred embodiment utilizes isocytosine and isoguanine in nucleic acids designed to be complementary to other probes, rather than target sequences, as this reduces non-specific hybridization, as is generally described in U.S. Patent No. 5,681,702. As used herein, the term "nucleoside" includes nucleotides as well as nucleoside and nucleotide analogs, and modified nucleosides such as amino modified nucleosides. In addition, "nucleoside" includes non-naturally occuring analog structures. Thus for example the individual units of a peptide nucleic acid, each containing a base, are referred to herein as a

nucleoside.

The present invention provides compositions and methods for identifying bases at target positions in a target nucleic acid. The term "target sequence" or "target nucleic acid" or grammatical equivalents herein means a nucleic acid sequence on a nucleic acid, generally a single strand of nucleic acid. The target sequence may be a portion of a gene, a regulatory sequence, genomic DNA, cDNA, RNA including mRNA and rRNA, or others. As is outlined herein, the target sequence may be a target sequence from a sample, or a secondary target such as a product of a reaction such as an amplification reaction, etc. It may be any length, with the understanding that longer sequences are more specific. As will be appreciated by those in the art, the complementary target sequence may take many forms. For example, it may be contained within a larger nucleic acid sequence, i.e. all or part of a gene or mRNA, a restriction fragment of a plasmid or genomic DNA, among others. As is outlined more fully below, probes are made to hybridize to target sequences to determine the presence or absence of the target sequence in a sample. Generally speaking, this term will be understood by those skilled in the art. The target sequence may also be comprised of different target domains; for example, a first target domain of the sample target sequence may hybridize to a capture probe or a portion of capture extender probe, a second target domain may hybridize to a portion of an amplifier probe, a label probe, or a different capture or capture extender probe, etc. The target domains may be adjacent or separated as indicated. Unless specified, the terms "first" and "second" are not meant to confer an orientation of the sequences with respect to the 5'-3' orientation of the target sequence. For example, assuming a 5'-3' orientation of the complementary target sequence, the first target domain may be located either 5' to the second domain, or 3' to the second domain.

As is more fully outlined below, the target sequence comprises positions for which sequence information is desired, generally referred to herein as the "target positions". In one embodiment, a single target position is elucidated; in a preferred embodiment, a plurality of target positions are elucidated. In general, the plurality of nucleotides in the target positions are contiguous with each other, although in some circumstances they may be separated by one or more nucleotides. By "plurality" as used herein is meant at least two. As used herein, the base which basepairs with the target position base in a hybrid is termed the "sequence position". That is, as more fully outlined below, the extension of a sequence primer results in nucleotides being added in the sequence positions, that are perfectly complementary to the nucleotides in the target positions. As will be appreciated by one of ordinary skill in the art, identification of a plurality of target positions in a target nucleotide sequence results in the determination of the nucleotide sequence of the target nucleotide sequence.

As will be appreciated by one of ordinary skill in the art, this system can take on a number of different

configurations, depending on the sequencing method used, the method of attaching a target sequence to a surface, etc. In general, the methods of the invention rely on the attachment of different target sequences to a solid support (which, as outlined below, can be accomplished in a variety of ways) to form an array. The target sequences comprise at least two domains: a first domain, for which sequence information is not desired, and to which a sequencing primer can hybridize, and a second domain, adjacent to the first domain, comprising the target positions for sequencing. A sequencing primer is hybridized to the target sequence, forming a hybridization complex, and then the sequencing primer is enzymatically extended by the addition of a first nucleotide into the first sequence position of the primer. This first nucleotide is then identified, as is outlined below, and then the process is repeated, to add nucleotides to the second, third, fourth, etc. sequence positions. The exact methods depend on the sequencing technique utilized, as outlined below.

Once the target sequence is associated onto the array as outlined below, the target sequence can be used in a variety of sequencing by synthesis reactions. These reactions are generally classified into several categories, outlined below.

SEQUENCING BY SYNTHESIS

As outlined herein, a number of sequencing by synthesis reactions are used to elucidate the identity of a plurality of bases at target positions within the target sequence. All of these reactions rely on the use of a target sequence comprising at least two domains; a first domain to which a sequencing primer will hybridize, and an adjacent second domain, for which sequence information is desired. Upon formation of the assay complex, extension enzymes are used to add dNTPs to the sequencing primer, and each addition of dNTP is "read" to determine the identity of the added dNTP. This may proceed for many cycles.

Pyrosequencing

In a preferred embodiment, pyrosequencing methods are done. Pyrosequencing is an extension method that can be used to add one or more nucleotides to the target positions. Pyrosequencing relies on the detection of a reaction product, pyrophosphate (PPi), produced during the addition of an NTP to a growing oligonucleotide chain, rather than on a label attached to the nucleotide. One molecule of PPi is produced per dNTP added to the extension primer. The detection of the PPi produced during the reaction is monitored using secondary enzymes; for example, preferred embodiments utilize secondary enzymes that convert the PPi into ATP, which also may be detected in a variety of ways, for example through a chemiluminescent reaction using luciferase and luciferin, or by the detection of NADPH. Thus, by running sequential reactions with each of the nucleotides, and monitoring the reaction products, the identity of the added base is determined.

Accordingly, the present invention provides methods of pyrosequencing on arrays; the arrays may be any number of different array configurations and substrates, as outlined herein, with microsphere arrays being particularly preferred. In this embodiment, the target sequence comprises a first domain that is substantially complementary to a sequencing primer, and an adjacent second domain that comprises a plurality of target positions. By "sequencing primer" herein is meant a nucleic acid that is substantially complementary to the first target domain, with perfect complementarity being preferred. As will be appreciated by those in the art, the length of the sequencing primer will vary with the conditions used. In general, the sequencing primer ranges from about 6 to about 500 or more basepairs in length, with from about 8 to about 100 being preferred, and from about 10 to about 25 being especially preferred.

Once the sequencing primer is added and hybridized to the target sequence to form a first hybridization complex (also sometimes referred to herein as an "assay complex"), the system is ready to initiate sequencing-by-synthesis. The methods described below make reference to the use of fiber optic bundle substrates with associated microspheres, but as will be appreciated by those in the art, any number of other substrates or solid supports may be used, or arrays that do not comprise microspheres.

The reaction is initiated by introducing the substrate comprising the hybridization complex comprising the target sequence (i.e. the array) to a solution comprising a first nucleotide, generally comprising deoxynucleoside-triphosphates (dNTPs). Generally, the dNTPs comprise dATP, dTTP, dCTP and dGTP. The nucleotides may be naturally occurring, such as deoxynucleotides, or non-naturally occurring, such as chain terminating nucleotides including dideoxynucleotides, as long as the enzymes used in the sequencing/detection reactions are still capable of recognizing the analogs. In addition, as more fully outlined below, for example in other sequencing-by-synthesis reactions, the nucleotides may comprise labels. The different dNTPs are added either to separate aliquots of the hybridization complex or preferably sequentially to the hybridization complex, as is more fully outlined below. In some embodiments it is important that the hybridization complex be exposed to a single type of dNTP at a time.

In addition, as will be appreciated by those in the art, the extension reactions of the present invention allow the precise incorporation of modified bases into a growing nucleic acid strand. Thus, any number of modified nucleotides may be incorporated for any number of reasons, including probing structure-function relationships (e.g. DNA:DNA or DNA:protein interactions), cleaving the nucleic acid, crosslinking the nucleic acid, incorporate mismatches, etc.

In addition to a first nucleotide, the solution also comprises an extension enzyme, generally a DNA

polymerase. Suitable DNA polymerases include, but are not limited to, the Klenow fragment of DNA polymerase I, SEQUENASE 1.0 and SEQUENASE 2.0 (U.S. Biochemical), T5 DNA polymerase and Phi29 DNA polymerase. If the dNTP is complementary to the base of the target sequence adjacent to the extension primer, the extension enzyme will add it to the extension primer, releasing pyrophosphate (PPi). Thus, the extension primer is modified, i.e. extended, to form a modified primer, sometimes referred to herein as a "newly synthesized strand". The incorporation of a dNTP into a newly synthesized nucleic acid strand releases PPi, one molecule of PPi per dNTP incorporated.

The release of pyrophosphate (PPi) during the DNA polymerase reaction can be quantitatively measured by many different methods and a number of enzymatic methods have been described; see Reeves et al., Anal. Biochem. 28:282 (1969); Guillory et al., Anal. Biochem. 39:170 (1971); Johnson et al., Anal. Biochem. 15:273 (1968); Cook et al., Anal. Biochem. 91:557 (1978); Drake et al., Anal. Biochem. 94:117 (1979); Ronaghi et al., Science 281:363 (1998); Barshop et al., Anal. Biochem. 197(1):266-272 (1991) WO93/23564; WO 98/28440; WO98/13523; Nyren et al., Anal. Biochem. 151:504 (1985); all of which are incorporated by reference. The latter method allows continuous monitoring of PPi and has been termed ELIDA (Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay). In a preferred embodiment, the PPi is detected utilizing UDP-glucose pyrophosphorylase, phosphoglucomutase and glucose 6-phosphate dehydrogenase. See Justesen, et al., Anal. Biochem. 207(1):90-93 (1992); Lust et al., Clin. Chem. Acta 66(2):241 (1976); and Johnson et al., Anal. Biochem. 26:137 (1968); all of which are hereby incorporated by reference. This reaction produces NADPH which can be detected fluoremetrically.
A preferred embodiment utilizes any method which can result in the generation of an optical signal, with preferred embodiments utilizing the generation of a chemiluminescent or fluorescent signal.

Generally, these methods rely on secondary enzymes to detect the PPi; these methods generally rely on enzymes that will convert PPi into ATP, which can then be detected. A preferred method monitors the creation of PPi by the conversion of PPi to ATP by the enzyme sulfurylase, and the subsequent production of visible light by firefly luciferase (see Ronaghi et al., supra, and Barshop, supra). In this method, the four deoxynucleotides (dATP, dGTP, dCTP and dTTP; collectively dNTPs) are added stepwise to a partial duplex comprising a sequencing primer hybridized to a single stranded DNA template and incubated with DNA polymerase, ATP sulfurylase (and its substrate, adenosine 5'-phosphosulphate (APS)) luciferase (and its substrate luciferin), and optionally a nucleotide-degrading enzyme such as apyrase. A dNTP is only incorporated into the growing DNA strand if it is complementary to the base in the template strand. The synthesis of DNA is accompanied by the release of PPi equal in molarity to the incorporated dNTP. The PPi is converted to ATP and the light generated by the luciferase is directly proportional to the amount of ATP. In some cases the unincorporated dNTPs and the produced ATP are degraded between each cycle by the nucleotide

degrading enzyme.

As will be appreciated by those in the art, if the target sequence comprises two or more of the same nucleotide in a row, more than one dNTP will be incorporated; however, the amount of PPi generated is directly proportional to the number of dNTPs incorporated and thus these sequences can be detected.

In addition, in a preferred embodiment, the dATP that is added to the reaction mixture is an analog that can be incorporated by the DNA polymerase into the growing oligonucleotide strand, but will not serve as a substrate for the second enzyme; for example, certain thiol-containing dATP analogs find particular use.

Accordingly, a preferred embodiment of the methods of the invention is as follows. A substrate comprising microspheres containing the target sequences and extension primers, forming hybridization complexes, is dipped or contacted with a volume (reaction chamber or well) comprising a single type of dNTP, an extension enzyme, and the reagents and enzymes necessary to detect PPi. If the dNTP is complementary to the base of the target portion of the target sequence adjacent to the extension primer, the dNTP is added, releasing PPi and generating detectable light, which is detected as generally described in U.S.S.N.s 09/151,877 and 09/189,543, and PCT US98/09163, all of which are hereby incorporated by reference. If the dNTP is not complementary, no detectable signal results. The substrate is then contacted with a second reaction chamber comprising a different dNTP and the additional components of the assay. This process is repeated to generate a readout of the sequence of the target sequence.

In a preferred embodiment, washing steps, i.e. the use of washing chambers, may be done in between the dNTP reaction chambers, as required. These washing chambers may optionally comprise a nucleotide-degrading enzyme, to remove any unreacted dNTP and decreasing the background signal, as is described in WO 98/28440, incorporated herein by reference. In a preferred embodiment a flow cell is used as a reaction chamber; following each reaction the unreacted dNTP is washed away and may be replaced with an additional dNTP to be examined.

As will be appreciated by those in the art, the system can be configured in a variety of ways, including both a linear progression or a circular one; for example, four substrates may be used that each can dip into one of four reaction chambers arrayed in a circular pattern. Each cycle of sequencing and reading is followed by a 90 degree rotation, so that each substrate then dips into the next reaction well. This allows a continuous series of sequencing reactions on multiple substrates in parallel.

In a preferred embodiment, one or more internal control sequences are used. That is, at least one microsphere in the array comprises a known sequence that can be used to verify that the reactions are proceeding correctly. In a preferred embodiment, at least four control sequences are used, each of which has a different nucleotide at each position: the first control sequence will have an adenosine at position 1, the second will have a cytosine, the third a guanosine, and the fourth a thymidine, thus ensuring that at least one control sequence is "lighting up" at each step to serve as an internal control.

In a preferred embodiment, the reaction is run for a number of cycles until the signal-to-noise ratio becomes low, generally from 20 to 70 cycles or more, with from about 30 to 50 being standard. In some embodiments, this is sufficient for the purposes of the experiment; for example, for the detection of certain mutations, including single nucleotide polymorphisms (SNPs), the experiment is designed such that the initial round of sequencing gives the desired information. In other embodiments, it is desirable to sequence longer targets, for example in excess of hundreds of bases. In this application, additional rounds of sequencing can be done.

For example, after a certain number of cycles, it is possible to stop the reaction, remove the newly synthesized strand using either a thermal step or a chemical wash, and start the reaction over, using for example the sequence information that was previously generated to make a new extension primer that will hybridize to the first target portion of the target sequence. That is, the sequence information generated in the first round is transferred to an oligonucleotide synthesizer, and a second extension primer is made for a second round of sequencing. In this way, multiple overlapping rounds of sequencing are used to generate long sequences from template nucleic acid strands. Alternatively, when a single target sequence contains a number of mutational "hot spots", primers can be generated using the known sequences in between these hot spots.

Additionally, the methods of the invention find use in the decoding of random microsphere arrays. That is, as described in U.S.S.N. 09/189,543, nucleic acids can be used as bead identifiers. By using sequencing-by-synthesis to read out the sequence of the nucleic acids, the beads can be decoded in a highly parallel fashion.

In addition, the methods find use in simultaneous analysis of multiple target sequence positions on a single array. For example, four separate sequence analysis reactions are performed. In the first reaction, positions containing a particular nucleotide ("A", for example) in the target sequence are analyzed. In three other reactions, C, G, and T are analyzed. An advantage of analyzing one base per reaction is that the baseline or background is flattened for the three bases excluded from the reaction. Therefore, the signal is more easily detected and the sensitivity of the assay is increased. Alternatively, each of the four sequencing reactions (A, G, C and T) can be performed simultaneously

with a nested set of primers providing a significant advantage in that primer synthesis can be made more efficient.

In another preferred embodiment each probe is represented by multiple beads in the array (see U.S.S.N. 09/287,573, filed April 6, 1999, hereby expressly incorporated by reference) . As a result, each experiment can be replicated many times in parallel. As outlined below, averaging the signal from each respective probe in an experiment also allows for improved signal to noise and increases the sensitivity of detecting subtle perturbations in signal intensity patterns. The use of redundancy and comparing the patterns obtained from two different samples (e.g. a reference and an unknown), results in highly paralleled and comparative sequence analysis that can be performed on complex nucleic acid samples.

As outlined herein, the pyrosequencing systems may be configured in a variety of ways; for example, the target sequence may be attached to the array (e.g. the beads) in a variety of ways, including the direct attachment of the target sequence to the array; the use of a capture probe with a separate extension probe; the use of a capture extender probe, a capture probe and a separate extension probe; the use of adapter sequences in the target sequence with capture and extension probes; and the use of a capture probe that also serves as the extension probe.

In addition, as will be appreciated by those in the art, the target sequence may comprise any number of sets of different first and second target domains; that is, depending on the number of target positions that may be elucidated at a time, there may be several "rounds" of sequencing occuring, each time using a different target domain.

One additional benefit of pyrosequencing for genotyping purposes is that since the reaction does not rely on the incorporation of labels into a growing chain, the unreacted extension primers need not be removed.

Thus, pyrosequencing kits and reactions require, in no particularly order, arrays comprising capture probes, sequencing primers, an extension enzyme, and secondary enzymes and reactants for the detection of PPi, generally comprising enzymes to convert PPi into ATP (or other NTPs), and enzymes and reactants to detect ATP.

Attachment of enzymes to arrays

In a preferred embodiment, particularly when secondary enzymes (i.e. enzymes other than extension enzymes) are used in the reaction, the enzyme(s) may be attached, preferably through the use of flexible linkers, to the sites on the array, e.g. the beads. For example, when pyrosequencing is done,

one embodiment utilizes detection based on the generation of a chemiluminescent signal in the "zone" around the bead. By attaching the secondary enzymes required to generate the signal, an increased concentration of the required enzymes is obtained in the immediate vicinity of the reaction, thus allowing for the use of less enzyme and faster reaction rates for detection. Thus, preferred embodiments utilize the attachment, preferably covalently (although as will be appreciated by those in the art, other attachment mechanisms may be used), of the non-extension secondary enzymes used to generate the signal. In some embodiments, the extension enzyme (e.g. the polymerase) may be attached as well, although this is not generally preferred.

The attachment of enzymes to array sites, particularly beads, is outlined in U.S.S.N. 09/287,573, hereby incorporated by reference, and will be appreciated by those in the art. In general, the use of flexible linkers are preferred, as this allows the enzymes to interact with the substrates. However, for some types of attachment, linkers are not needed. Attachment proceeds on the basis of the composition of the array site (i.e. either the substrate or the bead, depending on which array system is used) and the composition of the enzyme. In a preferred embodiment, depending on the composition of the array site (e.g. the bead), it will contain chemical functional groups for subsequent attachment of other moieties. For example, beads comprising a variety of chemical functional groups such as amines are commercially available. Preferred functional groups for attachment are amino groups, carboxy groups, oxo groups and thiol groups, with amino groups being particularly preferred. Using these functional groups, the enzymes can be attached using functional groups on the enzymes. For example, enzymes containing amino groups can be attached to particles comprising amino groups, for example using linkers as are known in the art; for example, homo-or hetero-bifunctional linkers as are well known (see 1994 Pierce Chemical Company catalog, technical section on cross-linkers, pages 155-200, incorporated herein by reference).

Reversible Chain Termination Methods

In a preferred embodiment, the sequencing-by-synthesis method utilized is reversible chain termination. In this embodiment, the rate of addition of dNTPs is controlled by using nucleotide analogs that contain a removable protecting group at the 3' position of the dNTP. The presence of the protecting group prevents further addition of dNTPs at the 3' end, thus allowing time for detection of the nucleotide added (for example, utilizing a labeled dNTP). After acquisition of the identity of the dNTP added, the protecting group is removed and the cycle repeated. In this way, dNTPs are added one at a time to the sequencing primer to allow elucidation of the nucleotides at the target positions. See U.S. Patent Nos. 5,902,723; 5,547,839; Metzker et al., Nucl. Acid Res. 22(20):4259 (1994); Canard et al., Gene 148(1):1-6 (1994); Dyatkina et al., Nucleic Acid Symp. Ser. 18:117-120 (1987); all of which are hereby expressly incorporated by reference.

Accordingly, the present invention provides methods and compositions for reversible chain termination sequencing-by-synthesis. Similar to pyrosequencing, the reaction requires the hybridization of a substantially complementary sequencing primer to a first target domain of a target sequence to form an assay complex.

The reaction is initiated by introducing the assay complex comprising the target sequence (i.e. the array) to a solution comprising a first nucleotide analog. By "nucleotide analog" in this context herein is meant a deoxynucleoside-triphosphate (also called deoxynucleotides or dNTPs, i.e. dATP, dTTP, dCTP and dGTP), that is further derivatized to be reversibly chain terminating. As will be appreciated by those in the art, any number of nucleotide analogs may be used, as long as a polymerase enzyme will still incorporate the nucleotide at the sequence position. Preferred embodiments utilize 3'-O-methyl-dNTPs (with photolytic removal of the protecting group), 3'-substituted-2'-dNTPs that contain anthranylic derivatives that are fluorescent (with alkali or enzymatic treatment for removal of the protecting group). The latter has the advantage that the protecting group is also the fluorescent label; upon cleavage, the label is also removed, which may serve to generally lower the background of the assay as well.

Again, the system may be configured and/or utilized in a number of ways. In a preferred embodiment, a set of nucleotide analogs such as derivatized dATP, derivatized dCTP, derivatized dGTP and derivatized dTTP is used, each with a different detectable and resolvable label, as outlined below. Thus, the identification of the base at the first sequencing position can be ascertained by the presence of the unique label.

Alternatively, a single label is used but the reactions are done sequentially. That is, the substrate comprising the array is first contacted with a reaction mixture of an extension enzyme and a single type of base with a first label, for example ddATP. The incorporation of the ddATP is monitored at each site on the array. The substrate is then contacted (with optional washing steps as needed) to a second reaction mixture comprising the extension enzyme and a second nucleotide, for example ddTTP. The reaction is then monitored; this can be repeated for each target position.

Once each reaction has been completed and the identification of the base at the sequencing position is ascertained, the terminating protecting group is removed, e.g. cleaved, leaving a free 3' end to repeat the sequence, using an extension enzyme to add a base to the 3' end of the sequencing primer when it is hybridized to the target sequence. As will be appreciated by those in the art, the cleavage conditions will vary with the protecting group chosen.

In a preferred embodiment, the nucleotide analogs comprise a detectable label. By "detection label" or

"detectable label" herein is meant a moiety that allows detection. This may be a primary label (directly detectable) or a secondary label (indirectly detectable).

In a preferred embodiment, the detection label is a primary label. A primary label is one that can be directly detected, such as a fluorophore. In general, primary labels fall into three classes: a) isotopic labels, which may be radioactive or heavy isotopes; b) magnetic, electrical, thermal labels; and c) colored or luminescent dyes. Labels can also include magnetic particles. Preferred labels include chromophores or phosphors but are preferably fluorescent dyes. Suitable dyes for use in the invention include, but are not limited to, fluorescent lanthanide complexes, including those of Europium and Terbium, fluorescein, rhodamine, tetramethylrhodamine, eosin, erythrosin, coumarin, methyl-coumarins, pyrene, Malacite green, stilbene, Lucifer Yellow, Cascade Blue™, Texas Red, phycoerythrin, Cy dyes, Bodipy, Alexa dyes, so called "quantum dots" (also referred to in the literature as "nanocrystals") and others described in the 6th Edition of the Molecular Probes Handbook by Richard P. Haugland, hereby expressly incorporated by reference.

In a preferred embodiment, the detection label is a secondary label. A secondary label is one that is indirectly detected. This may include, but is not limited to, secondary labels that a) bind or react with a primary label for detection; or b) interact with secondary moieties to produce a label (e.g. enzymes and flurogenic or chromogenic substrates).

In a preferred embodiment, the secondary label is a binding partner pair. For example, the label may be a hapten or antigen, which will bind its binding partner that comprises a primary label. For example, suitable binding partner pairs include, but are not limited to: antigens (such as proteins (including peptides)) and antibodies (including fragments thereof (FAbs, etc.)); proteins and small molecules, including biotin/streptavidin and digoxygenin and antibodies; enzymes and substrates or inhibitors; other protein-protein interacting pairs; receptor-ligands; and carbohydrates and their binding partners, are also suitable binding pairs. Nucleic acid - nucleic acid binding proteins pairs are also useful. In general, the smaller of the pair is attached to the NTP (or the probe) for incorporation into the extension primer.

In a preferred embodiment, the binding partner pair comprises biotin or imino-biotin and streptavidin. Imino-biotin is particularly preferred when the methods require the later separation of the pair, as imino-biotin disassociates from streptavidin in pH 4.0 buffer while biotin requires harsh denaturants (e.g. 6 M guanidinium HCl, pH 1.5 or 90% formamide at 95°C).

In a preferred embodiment, the binding partner pair comprises a primary detection label (attached to the NTP and therefore to the extended primer) and an antibody that will specifically bind to the primary

detection label. By "specifically bind" herein is meant that the partners bind with specificity sufficient to differentiate between the pair and other components or contaminants of the system. The binding should be sufficient to remain bound under the conditions of the assay, including wash steps to remove non-specific binding. In some embodiments, the dissociation constants of the pair will be less than about 10^{-4}-10^{-6} M^{-1}, with less than about 10^{-5} to 10^{-9} M^{-1} being preferred and less than about 10^{-7} - 10^{-9} M^{-1} being particularly preferred.

In addition to a first nucleotide, the solution also comprises an extension enzyme, generally a DNA polymerase, as outlined above for pyrosequencing.

In a preferred embodiment, the protecting group also comprises a label. That is, as outlined in Canard et al., supra, the protecting group can serve as either a primary or secondary label, with the former being preferred. This is particularly preferred as the removal of the label at each round results in less background noise, less quenching and less crosstalk.

In this way, reversible chain termination sequencing is accomplished.

Time-resolved sequencing

In a preferred embodiment, time-resolved sequencing is done. This embodiment relies on controlling the reaction rate of the extension reaction and/or using a fast imaging system. Basically, the method involves a simple extension reaction that is either "slowed down", or imaged using a fast system, or both. What is important is that the rate of polymerization (extension) is significantly slower than the rate of image capture.

To allow for real time sequencing, parameters such as the speed of the detector (millisecond speed is preferred), and rate of polymerization will be controlled such that the rate of polymerization is significantly slower than the rate of image capture. Polymerization rates on the order of kilobases per minute (e.g. ~10 milliseconds/nucleotide), which can be adjusted, should allow a sufficiently wide window to find conditions where the sequential addition of two nucleotides can be resolved. The DNA polymerization reaction, which has been studied intensively, can easily be reconstituted in vitro and controlled by varying a number of parameters including reaction temperature and the concentration of nucleotide triphosphates.

In addition, the polymerase can be applied to the primer-template complex prior to initiating the reaction. This serves to synchronize the reaction. Numerous polymerases are available. Some examples include, but are not limited to polymerases with 3' to 5' exonuclease activity, other nuclease activities, polymerases with different processivity, affinities for modified and unmodified nucleotide

triphosphates, temperature optima, stability, and the like.

Thus, in this embodiment, the reaction proceeds as outlined above. The target sequence, comprising a first domain that will hybridize to a sequencing primer and a second domain comprising a plurality of target positions, is attached to an array as outlined below. The sequencing primers are added, along with an extension enzyme, as outlined herein, and dNTPs are added. Again, as outlined above, either four differently labeled dNTPs may be used simultaneously or, four different sequential reactions with a single label are done. In general, the dNTPs comprise either a primary or a secondary label, as outlined above.

In a preferred embodiment, the extension enzyme is one that is relatively "slow". This may be accomplished in several ways. In one embodiment, polymerase variants are used that have a lower polymerization rate than wild-type enzymes. Alternatively, the reaction rate may be controlled by varying the temperature and the concentration of dNTPs.

In a preferred embodiment, a fast (millisecond) high-sensitivity imaging system is used.

In one embodiment, DNA polymerization (extension) is monitored using light scattering, as is outlined in Johnson et al., Anal. Biochem. 136(1):192 (1984), hereby expressly incorporated by reference.

ATTACHMENT OF TARGET SEQUENCES TO ARRAYS

As is generally described herein, there are a variety of methods that can be used to attach target sequences to the solid supports of the invention, particularly to the microspheres that are distributed on a surface of a substrate. Most of these methods generally rely on capture probes attached to the array. However, the attachment may be direct or indirect. Direct attachment includes those situations wherein an endogeneous portion of the target sequence hybridizes to the capture probe, or where the target sequence has been manipulated to contain exogeneous adapter sequences that are added to the target sequence, for example during an amplification reaction. Alternatively, the target sequences may be directly attached to the beads. Indirect attachment utilizes one or more secondary probes, termed a "capture extender probe". These methods are further described in "Addressing Arrays using Sequence Specific Adapters", filed October 22, 1999, no U.S.S.N. received yet, herein incorporated by reference.

In a preferred embodiment, direct attachment is done, as is generally depicted in Figure 1A. In this embodiment, the target sequence comprises a first target domain that hybridizes to all or part of the capture probe.

In a preferred embodiment, direct attachment is accomplished through the use of adapter sequences. An "adapter sequence" as used herein is a sequence that is generally not native to the target sequence, i.e. is exogeneous, but is added during an amplification reaction, such as PCR or any of the other amplification techniques. In this embodiment, one or more of the amplification primers comprises a first portion comprising the adapter sequence and a second portion comprising the primer sequence. Extending the amplification primer as is well known in the art results in target sequences that comprise the adapter sequences. The adapter sequences are designed to be substantially complementary to capture probes.

In a preferred embodiment, indirect attachment of the target sequence to the array is done, through the use of capture extender probes. "Capture extender" probes are generally depicted in Figure 1C, and other figures, and have a first portion that will hybridize to all or part of the capture probe, and a second portion that will hybridize to a first portion of the target sequence. Two capture extender probes may also be used. This has generally been done to stabilize assay complexes for example when the target sequence is large, or when large amplifier probes (particularly branched or dendrimer amplifier probes) are used.

When only capture probes are utilized, it is necessary to have unique capture probes for each target sequence; that is, the surface must be customized to contain unique capture probes; e.g. each bead comprises a different capture probe. Only a single type of capture probe should be bound to a bead; however, different beads should contain different capture probes so that different target sequences bind to different beads.

Alternatively, the use of adapter sequences and capture extender probes allow the creation of more "universal" surfaces. In a preferred embodiment, an array of different and usually artificial capture probes are made; that is, the capture probes do not have complementarity to known target sequences. The adapter sequences can then be added to any target sequences, or soluble capture extender probes are made; this allows the manufacture of only one kind of array, with the user able to customize the array through the use of adapter sequences or capture extender probes. This then allows the generation of customized soluble probes, which as will be appreciated by those in the art is generally simpler and less costly.

As will be appreciated by those in the art, the length of the adapter sequences will vary, depending on the desired "strength" of binding and the number of different adapters desired. In a preferred embodiment, adapter sequences range from about 6 to about 500 basepairs in length, with from about 8 to about 100 being preferred, and from about 10 to about 25 being particularly preferred.

In one embodiment, microsphere arrays containing a single type of capture probe are made; in this embodiment, the capture extender probes are added to the beads prior to loading on the array. The capture extender probes may be additionally fixed or crosslinked, as necessary.

In a preferred embodiment, as outlined in Figure 1B, the capture probe comprises the sequencing primer; that is, after hybridization to the target sequence, it is the capture probe itself that is extended during the synthesis reaction.

In one embodiment, capture probes are not used, and the target sequences are attached directly to the sites on the array. For example, libraries of clonal nucleic acids, including DNA and RNA, are used. In this embodiment, individual nucleic acids are prepared, generally using conventional methods (including, but not limited to, propagation in plasmid or phage vectors, amplification techniques including PCR, etc.). The nucleic acids are preferably arrayed in some format, such as a microtiter plate format, and either spotted or beads are added for attachment of the libraries.

Attachment of the clonal libraries (or any of the nucleic acids outlined herein) may be done in a variety of ways, as will be appreciated by those in the art, including, but not limited to, chemical or affinity capture (for example, including the incorporation of derivatized nucleotides such as AminoLink or biotinylated nucleotides that can then be used to attach the nucleic acid to a surface, as well as affinity capture by hybridization), cross-linking, and electrostatic attachment, etc.

In a preferred embodiment, affinity capture is used to attach the clonal nucleic acids to the surface. For example, cloned nucleic acids can be derivatized, for example with one member of a binding pair, and the beads derivatized with the other member of a binding pair. Suitable binding pairs are as described herein for secondary labels and IBL/DBL pairs. For example, the cloned nucleic acids may be biotinylated (for example using enzymatic incorporate of biotinylated nucleotides, for by photoactivated cross-linking of biotin). Biotinylated nucleic acids can then be captured on streptavidin-coated beads, as is known in the art. Similarly, other hapten-receptor combinations can be used, such as digoxigenin and anti-digoxigenin antibodies. Alternatively, chemical groups can be added in the form of derivatized nucleotides, that can them be used to add the nucleic acid to the surface.

Preferred attachments are covalent, although even relatively weak interactions (i.e. non-covalent) can be sufficient to attach a nucleic acid to a surface, if there are multiple sites of attachment per each nucleic acid. Thus, for example, electrostatic interactions can be used for attachment, for example by having beads carrying the opposite charge to the bioactive agent.

Similarly, affinity capture utilizing hybridization can be used to attach cloned nucleic acids to beads.

For example, as is known in the art, polyA+RNA is routinely captured by hybridization to oligo-dT beads; this may include oligo-dT capture followed by a cross-linking step, such as psoralen crosslinking). If the nucleic acids of interest do not contain a polyA tract, one can be attached by polymerization with terminal transferase, or via ligation of an oligoA linker, as is known in the art.

Alternatively, chemical crosslinking may be done, for example by photoactivated crosslinking of thymidine to reactive groups, as is known in the art.

In general, special methods are required to decode clonal arrays, as is more fully outlined below.

All of the methods and compositions herein are drawn to methods of sequencing target nucleic acids at the target positions. These reactions are generally detected on arrays, and particularly microsphere arrays, as is outlined herein.

ARRAYS

The present invention provides array compositions comprising at least a first substrate with a surface comprising individual sites. By "array" or "biochip" herein is meant a plurality of nucleic acids in an array format; the size of the array will depend on the composition and end use of the array. Nucleic acids arrays are known in the art, and can be classified in a number of ways; both patterned arrays (e.g. the ability to resolve chemistries at discrete sites), and random arrays are included. Ordered arrays include, but are not limited to, those made using photolithography techniques (Affymetrix GeneChip™), spotting techniques (Synteni and others), printing techniques (Hewlett Packard and Rosetta), three dimensional "gel pad" arrays, etc. A preferred embodiment utilizes microspheres on a variety of substrates including fiber optic bundles, as are outlined in PCTs US98/21193, PCT US99/14387 and PCT US98/05025; WO98/50782; and U.S.S.N.s 09/287,573, 09/151,877, 09/256,943, 09/316,154, 60/119,323, 09/315,584; all of which are expressly incorporated by reference. While much of the discussion below is directed to the use of microsphere arrays on fiber optic bundles, any array format of nucleic acids on solid supports may be utilized.

The present invention provides array compositions comprising substrates with surfaces comprising discrete sites. By "array" or "biochip" herein is meant a plurality of target analyte sets in an array format; the size of the array will depend on the composition and end use of the array. That is, each site on the array comprises a set of target analytes. Nucleic acids arrays are known in the art, and can be classified in a number of ways; both ordered arrays (e.g. the ability to resolve chemistries at discrete sites), and random arrays are included. Ordered arrays include, but are not limited to, those made using photolithography techniques (Affymetrix GeneChip™), spotting techniques (Synteni and

others), printing techniques (Hewlett Packard and Rosetta), three dimensional "gel pad" arrays, etc. A preferred embodiment utilizes microspheres on a variety of substrates including fiber optic bundles, as are outlined in PCTs US98/21193, PCT US99/14387 and PCT US98/05025; WO98/50782; and U.S.S.N.s 09/287,573, 09/151,877, 09/256,943, 09/316,154, 60/119,323, 09/315,584; all of which are expressly incorporated by reference. While much of the discussion below is directed to the use of microsphere arrays on substrates such as fiber optic bundles, any array format of nucleic acids on solid supports may be utilized.

Arrays containing from about 2 different nucleic acids (e.g. different beads, when beads are used) to many millions can be made, with very large fiber optic arrays being possible. Generally, the array will comprise from two to as many as a billion or more, depending on the size of the beads and the substrate, as well as the end use of the array, thus very high density, high density, moderate density, low density and very low density arrays may be made. Preferred ranges for very high density arrays are from about 10,000,000 to about 2,000,000,000, with from about 100,000,000 to about 1,000,000,000 being preferred (all numbers being in square cm). High density arrays range about 100,000 to about 10,000,000, with from about 1,000,000 to about 5,000,000 being particularly preferred. Moderate density arrays range from about 10,000 to about 100,000 being particularly preferred, and from about 20,000 to about 50,000 being especially preferred. Low density arrays are generally less than 10,000, with from about 1,000 to about 5,000 being preferred. Very low density arrays are less than 1,000, with from about 10 to about 1000 being preferred, and from about 100 to about 500 being particularly preferred. In some embodiments, the compositions of the invention may not be in array format; that is, for some embodiments, compositions comprising a single bioactive agent may be made as well. In addition, in some arrays, multiple substrates may be used, either of different or identical compositions. Thus for example, large arrays may comprise a plurality of smaller substrates.

In addition, one advantage of the present compositions is that particularly through the use of fiber optic technology, extremely high density arrays can be made. Thus for example, because beads of 200 μm or less (with beads of 200 nm possible) can be used, and very small fibers are known, it is possible to have as many as 40,000 or more (in some instances, 1 million) different elements (e.g. fibers and beads) in a 1 mm^2 fiber optic bundle, with densities of greater than 25,000,000 individual beads and fibers (again, in some instances as many as 50-100 million) per 0.5 cm^2 obtainable (4 million per square cm for 5 μ center-to-center and 100 million per square cm for 1 μ center-to-center).

By "substrate" or "solid support" or other grammatical equivalents herein is meant any material that can be modified to contain discrete individual sites appropriate for the attachment or association of

beads and is amenable to at least one detection method. As will be appreciated by those in the art, the number of possible substrates is very large. Possible substrates include, but are not limited to, glass and modified or functionalized glass, plastics (including acrylics, polystyrene and copolymers of styrene and other materials, polypropylene, polyethylene, polybutylene, polyurethanes, Teflon, etc.), polysaccharides, nylon or nitrocellulose, resins, silica or silica-based materials including silicon and modified silicon, carbon, metals, inorganic glasses, plastics, optical fiber bundles, and a variety of other polymers. In general, the substrates allow optical detection and do not themselves appreciably fluoresce.

Generally the substrate is flat (planar), although as will be appreciated by those in the art, other configurations of substrates may be used as well; for example, three dimensional configurations can be used, for example by embedding the beads in a porous block of plastic that allows sample access to the beads and using a confocal microscope for detection. Similarly, the beads may be placed on the inside surface of a tube, for flow-through sample analysis to minimize sample volume. Preferred substrates include optical fiber bundles as discussed below, and flat planar substrates such as glass, polystyrene and other plastics and acrylics.

In a preferred embodiment, the substrate is an optical fiber bundle or array, as is generally described in U.S.S.N.s 08/944,850 and 08/519,062, PCT US98/05025, and PCT US98/09163, all of which are expressly incorporated herein by reference. Preferred embodiments utilize preformed unitary fiber optic arrays. By "preformed unitary fiber optic array" herein is meant an array of discrete individual fiber optic strands that are co-axially disposed and joined along their lengths. The fiber strands are generally individually clad. However, one thing that distinguished a preformed unitary array from other fiber optic formats is that the fibers are not individually physically manipulatable; that is, one strand generally cannot be physically separated at any point along its length from another fiber strand.

At least one surface of the substrate is modified to contain discrete, individual sites for later association of microspheres. These sites may comprise physically altered sites, i.e. physical configurations such as wells or small depressions in the substrate that can retain the beads, such that a microsphere can rest in the well, or the use of other forces (magnetic or compressive), or chemically altered or active sites, such as chemically functionalized sites, electrostatically altered sites, hydrophobically/ hydrophilically functionalized sites, spots of adhesive, etc.

The sites may be a pattern, i.e. a regular design or configuration, or randomly distributed. A preferred embodiment utilizes a regular pattern of sites such that the sites may be addressed in the X-Y coordinate plane. "Pattern" in this sense includes a repeating unit cell, preferably one that allows a high density of beads on the substrate. However, it should be noted that these sites may not be

discrete sites. That is, it is possible to use a uniform surface of adhesive or chemical functionalities, for example, that allows the attachment of beads at any position. That is, the surface of the substrate is modified to allow attachment of the microspheres at individual sites, whether or not those sites are contiguous or non-contiguous with other sites. Thus, the surface of the substrate may be modified such that discrete sites are formed that can only have a single associated bead, or alternatively, the surface of the substrate is modified and beads may go down anywhere, but they end up at discrete sites.

In a preferred embodiment, the surface of the substrate is modified to contain wells, i.e. depressions in the surface of the substrate. This may be done as is generally known in the art using a variety of techniques, including, but not limited to, photolithography, stamping techniques, molding techniques and microetching techniques. As will be appreciated by those in the art, the technique used will depend on the composition and shape of the substrate.

In a preferred embodiment, physical alterations are made in a surface of the substrate to produce the sites. In a preferred embodiment, the substrate is a fiber optic bundle and the surface of the substrate is a terminal end of the fiber bundle, as is generally described in 08/818,199 and 09/151,877, both of which are hereby expressly incorporated by reference. In this embodiment, wells are made in a terminal or distal end of a fiber optic bundle comprising individual fibers. In this embodiment, the cores of the individual fibers are etched, with respect to the cladding, such that small wells or depressions are formed at one end of the fibers. The required depth of the wells will depend on the size of the beads to be added to the wells.

Generally in this embodiment, the microspheres are non-covalently associated in the wells, although the wells may additionally be chemically functionalized as is generally described below, cross-linking agents may be used, or a physical barrier may be used, i.e. a film or membrane over the beads.

In a preferred embodiment, the surface of the substrate is modified to contain chemically modified sites, that can be used to attach, either covalently or non-covalently, the microspheres of the invention to the discrete sites or locations on the substrate. "Chemically modified sites" in this context includes, but is not limited to, the addition of a pattern of chemical functional groups including amino groups, carboxy groups, oxo groups and thiol groups, that can be used to covalently attach microspheres, which generally also contain corresponding reactive functional groups; the addition of a pattern of adhesive that can be used to bind the microspheres (either by prior chemical functionalization for the addition of the adhesive or direct addition of the adhesive); the addition of a pattern of charged groups (similar to the chemical functionalities) for the electrostatic attachment of the microspheres, i.e. when the microspheres comprise charged groups opposite to the sites; the addition of a pattern of chemical

functional groups that renders the sites differentially hydrophobic or hydrophilic, such that the addition of similarly hydrophobic or hydrophilic microspheres under suitable experimental conditions will result in association of the microspheres to the sites on the basis of hydroaffinity. For example, the use of hydrophobic sites with hydrophobic beads, in an aqueous system, drives the association of the beads preferentially onto the sites. As outlined above, "pattern" in this sense includes the use of a uniform treatment of the surface to allow attachment of the beads at discrete sites, as well as treatment of the surface resulting in discrete sites. As will be appreciated by those in the art, this may be accomplished in a variety of ways.

In a preferred embodiment, the compositions of the invention further comprise a population of microspheres. By "population" herein is meant a plurality of beads as outlined above for arrays. Within the population are separate subpopulations, which can be a single microsphere or multiple identical microspheres. That is, in some embodiments, as is more fully outlined below, the array may contain only a single bead for each capture probe; preferred embodiments utilize a plurality of beads of each type.

By "microspheres" or "beads" or "particles" or grammatical equivalents herein is meant small discrete particles. The composition of the beads will vary, depending on the class of capture probe and the method of synthesis. Suitable bead compositions include those used in peptide, nucleic acid and organic moiety synthesis, including, but not limited to, plastics, ceramics, glass, polystyrene, methylstyrene, acrylic polymers, paramagnetic materials, thoria sol, carbon graphite, titanium dioxide, latex or cross-linked dextrans such as Sepharose, cellulose, nylon, cross-linked micelles and Teflon may all be used. *"Microsphere Detection Guide"* from Bangs Laboratories, Fishers IN is a helpful guide.

The beads need not be spherical; irregular particles may be used. In addition, the beads may be porous, thus increasing the surface area of the bead available for either capture probe attachment or tag attachment. The bead sizes range from nanometers, i.e. 100 nm, to millimeters, i.e. 1 mm, with beads from about 0.2 micron to about 200 microns being preferred, and from about 0.5 to about 5 micron being particularly preferred, although in some embodiments smaller beads may be used.

It should be noted that a key component of the invention is the use of a substrate/bead pairing that allows the association or attachment of the beads at discrete sites on the surface of the substrate, such that the beads do not move during the course of the assay.

Each element of the array (e.g. each bead) comprises a capture probe, although as will be appreciated by those in the art, there may be some microspheres which do not contain a capture

probe, depending on the synthetic methods.

In a preferred embodiment, each site on the array, e.g. each microsphere when microsphere arrays are utilized, comprises a capture probe. By "capture probe" or "capture nucleic acid" or "anchor probe" herein is meant a component of an assay complex as defined herein that allows the attachment of a target sequence to the substrate for the purposes of detection. As is more fully outlined below, attachment of the target sequence to the capture probe may be direct (i.e. the target sequence hybridizes to the capture probe) or indirect (one or more adapter probes are used). In a preferred embodiment, the capture probes are covalently attached to the microspheres. By "covalently attached" herein is meant that two moieties are attached by at least one bond, including sigma bonds, pi bonds and coordination bonds. In addition, as is more fully outlined below, the capture probes may have both nucleic and non-nucleic acid portions. Thus, for example, flexible linkers such as alkyl groups, may be used.

In general, probes of the present invention are designed to be complementary to a target sequence (either the target analyte sequence of the sample or to other probe sequences, such as the product of an amplification reaction or an adapter sequences, as is described herein), such that hybridization of the target and the probes of the present invention occurs. This complementarily need not be perfect; there may be any number of base pair mismatches that will interfere with hybridization between the target sequence and the single stranded nucleic acids of the present invention. However, if the number of mutations is so great that no hybridization can occur under even the least stringent of hybridization conditions, the sequence is not a complementary target sequence. Thus, by "substantially complementary" herein is meant that the probes are sufficiently complementary to the target sequences to hybridize under the selected reaction conditions. High stringency conditions are known in the art; see for example Maniatis et al., Molecular Cloning: A Laboratory Manual, 2d Edition, 1989, and Short Protocols in Molecular Biology, ed. Ausubel, et al., both of which are hereby incorporated by reference. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes, "Overview of principles of hybridization and the strategy of nucleic acid assays" (1993). Generally, stringent conditions are selected to be about 5-10°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength pH. The T_m is the temperature (under defined ionic strength, pH and nucleic acid concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at T_m, 50% of the probes are occupied at equilibrium). Stringent conditions will be those in which the salt concentration is less than about 1.0 M sodium ion, typically about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH

7.0 to 8.3 and the temperature is at least about 30°C for short probes (e.g. 10 to 50 nucleotides) and at least about 60°C for long probes (e.g. greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. In another embodiment, less stringent hybridization conditions are used; for example, moderate or low stringency conditions may be used, as are known in the art; see Maniatis and Ausubel, supra, and Tijssen, supra.

In general, as is known in the art, the length of the capture probes used to attach the target sequences to the array may vary. Preferred embodiments utilize probes ranging from about 6 to about 500 bases, with from about 8 to about 100 being preferred, and from about 10 to about 25 being particularly preferred.

In a preferred embodiment, capture probes are used to attach the target sequences to the substrate. This may be done in a variety of ways, three of which are depicted in Figure 1A and described in detail above. In one embodiment, the capture probe hybridizes to one domain of the target sequence, and the sequencing primer hybridizes to another domain of the target sequence, which may be either an endogeneous sequence or an adapter sequence. In a further embodiment, the capture probe hybridizes to a first domain of a capture extender probe (also referred to herein as an adapter probe), and a second domain of the capture extender probe hybridizes to a first domain of the target sequence. The sequencing primer hybridizes to a second domain of the target sequence. In an alternative embodiment, the capture probe serves as the sequencing primer, described below. Finally, in some embodiments, as outlined below, no capture probes are used and the target sequences themselves are directly attached to the arrays, e.g. the microspheres when they are used.

Attachment of the probe (or target) nucleic acids may be done in a variety of ways, as will be appreciated by those in the art, including, but not limited to, chemical or affinity capture (for example, including the incorporation of derivatized nucleotides such as AminoLink or biotinylated nucleotides that can then be used to attach the nucleic acid to a surface, as well as affinity capture by hybridization), cross-linking, and electrostatic attachment, etc. In a preferred embodiment, affinity capture is used to attach the nucleic acids to the beads. For example, nucleic acids can be derivatized, for example with one member of a binding pair, and the beads derivatized with the other member of a binding pair. Suitable binding pairs are as described herein for IBL/DBL pairs. For example, the nucleic acids may be biotinylated (for example using enzymatic incorporate of biotinylated nucleotides, for by photoactivated cross-linking of biotin). Biotinylated nucleic acids can then be captured on streptavidin-coated beads, as is known in the art. Similarly, other hapten-receptor combinations can be used, such as digoxigenin and anti-digoxigenin antibodies. Alternatively, chemical groups can be added in the form of derivatized nucleotides, that can them be used to add the nucleic acid to the surface.

Preferred attachments are covalent, although even relatively weak interactions (i.e. non-covalent) can be sufficient to attach a nucleic acid to a surface, if there are multiple sites of attachment per each nucleic acid. Thus, for example, electrostatic interactions can be used for attachment, for example by having beads carrying the opposite charge to the bioactive agent.

Similarly, affinity capture utilizing hybridization can be used to attach nucleic acids to beads. For example, as is known in the art, polyA+RNA is routinely captured by hybridization to oligo-dT beads; this may include oligo-dT capture followed by a cross-linking step, such as psoralen crosslinking). If the nucleic acids of interest do not contain a polyA tract, one can be attached by polymerization with terminal transferase, or via ligation of an oligoA linker, as is known in the art.

Alternatively, chemical crosslinking may be done, for example by photoactivated crosslinking of thymidine to reactive groups, as is known in the art.

In a preferred embodiment, each element of the array, e.g. each bead, comprises a single type of capture probe, although a plurality of individual capture probes are preferably attached to each bead. Similarly, preferred embodiments utilize more than one microsphere containing a unique capture probe; that is, there is redundancy built into the system by the use of subpopulations of microspheres, each microsphere in the subpopulation containing the same capture probe.

As will be appreciated by those in the art, the capture probes may either be synthesized directly on the beads, or they may be made and then attached after synthesis. In a preferred embodiment, linkers are used to attach the capture probes to the beads, to allow both good attachment, sufficient flexibility to allow good interaction with the target molecule, and to avoid undesirable binding reactions.

In a preferred embodiment, the capture probes are synthesized directly on the beads. As is known in the art, many classes of chemical compounds are currently synthesized on solid supports, such as peptides, organic moieties, and nucleic acids. It is a relatively straightforward matter to adjust the current synthetic techniques to use beads.

In a preferred embodiment, the capture probes are synthesized first, and then covalently attached to the beads. As will be appreciated by those in the art, this will be done depending on the composition of the capture probes and the beads. The functionalization of solid support surfaces such as certain polymers with chemically reactive groups such as thiols, amines, carboxyls, etc. is generally known in the art. Accordingly, "blank" microspheres may be used that have surface chemistries that facilitate the attachment of the desired functionality by the user. Some examples of these surface chemistries for blank microspheres include, but are not limited to, amino groups including aliphatic and aromatic

amines, carboxylic acids, aldehydes, amides, chloromethyl groups, hydrazide, hydroxyl groups, sulfonates and sulfates.

When microsphere arrays are used, an encoding/decoding system must be used. That is, since the beads are generally put onto the substrate randomly, there are several ways to correlate the functionality on the bead with its location, including the incorporation of unique optical signatures, generally fluorescent dyes, that could be used to identify the chemical functionality on any particular bead. This allows the synthesis of the candidate agents (i.e. compounds such as nucleic acids and antibodies) to be divorced from their placement on an array, i.e. the candidate agents may be synthesized on the beads, and then the beads are randomly distributed on a patterned surface. Since the beads are first coded with an optical signature, this means that the array can later be "decoded", i.e. after the array is made, a correlation of the location of an individual site on the array with the bead or candidate agent at that particular site can be made. This means that the beads may be randomly distributed on the array, a fast and inexpensive process as compared to either the in situ synthesis or spotting techniques of the prior art.

However, the drawback to these methods is that for a large array, the system requires a large number of different optical signatures, which may be difficult or time-consuming to utilize. Accordingly, the present invention provides several improvements over these methods, generally directed to methods of coding and decoding the arrays. That is, as will be appreciated by those in the art, the placement of the capture probes is generally random, and thus a coding/decoding system is required to identify the probe at each location in the array. This may be done in a variety of ways, as is more fully outlined below, and generally includes: a) the use a decoding binding ligand (DBL), generally directly labeled, that binds to either the capture probe or to identifier binding ligands (IBLs) attached to the beads; b) positional decoding, for example by either targeting the placement of beads (for example by using photoactivatible or photocleavable moieties to allow the selective addition of beads to particular locations), or by using either sub-bundles or selective loading of the sites, as are more fully outlined below; c) selective decoding, wherein only those beads that bind to a target are decoded; or d) combinations of any of these. In some cases, as is more fully outlined below, this decoding may occur for all the beads, or only for those that bind a particular target sequence. Similarly, this may occur either prior to or after addition of a target sequence. In addition, as outlined herein, the target sequences detected may be either a primary target sequence (e.g. a patient sample), or a reaction product from one of the methods described herein (e.g. an extended SBE probe, a ligated probe, a cleaved signal probe, etc.).

Once the identity (i.e. the actual agent) and location of each microsphere in the array has been fixed, the array is exposed to samples containing the target sequences, although as outlined below, this can

be done prior to or during the analysis as well. The target sequences can hybridize (either directly or indirectly) to the capture probes as is more fully outlined below, and results in a change in the optical signal of a particular bead.

In the present invention, "decoding" does not rely on the use of optical signatures, but rather on the use of decoding binding ligands that are added during a decoding step. The decoding binding ligands will bind either to a distinct identifier binding ligand partner that is placed on the beads, or to the capture probe itself. The decoding binding ligands are either directly or indirectly labeled, and thus decoding occurs by detecting the presence of the label. By using pools of decoding binding ligands in a sequential fashion, it is possible to greatly minimize the number of required decoding steps.

In some embodiments, the microspheres may additionally comprise identifier binding ligands for use in certain decoding systems. By "identifier binding ligands" or "IBLs" herein is meant a compound that will specifically bind a corresponding decoder binding ligand (DBL) to facilitate the elucidation of the identity of the capture probe attached to the bead. That is, the IBL and the corresponding DBL form a binding partner pair. By "specifically bind" herein is meant that the IBL binds its DBL with specificity sufficient to differentiate between the corresponding DBL and other DBLs (that is, DBLs for other IBLs), or other components or contaminants of the system. The binding should be sufficient to remain bound under the conditions of the decoding step, including wash steps to remove non-specific binding. In some embodiments, for example when the IBLs and corresponding DBLs are proteins or nucleic acids, the dissociation constants of the IBL to its DBL will be less than about 10^{-4}-10^{-6} M^{-1}, with less than about 10^{-5} to 10^{-9} M^{-1} being preferred and less than about 10^{-7} -10^{-9} M^{-1} being particularly preferred.

IBL-DBL binding pairs are known or can be readily found using known techniques. For example, when the IBL is a protein, the DBLs include proteins (particularly including antibodies or fragments thereof (FAbs, etc.)) or small molecules, or vice versa (the IBL is an antibody and the DBL is a protein). Metal ion- metal ion ligands or chelators pairs are also useful. Antigen-antibody pairs, enzymes and substrates or inhibitors, other protein-protein interacting pairs, receptor-ligands, complementary nucleic acids, and carbohydrates and their binding partners are also suitable binding pairs. Nucleic acid - nucleic acid binding proteins pairs are also useful. Similarly, as is generally described in U.S. Patents 5,270,163, 5,475,096, 5,567,588, 5,595,877, 5,637,459, 5,683,867,5,705,337, and related patents, hereby incorporated by reference, nucleic acid "aptamers" can be developed for binding to virtually any target; such an aptamer-target pair can be used as the IBL-DBL pair. Similarly, there is a wide body of literature relating to the development of binding pairs based on combinatorial chemistry methods.

In a preferred embodiment, the IBL is a molecule whose color or luminescence properties change in the presence of a selectively-binding DBL. For example, the IBL may be a fluorescent pH indicator whose emission intensity changes with pH. Similarly, the IBL may be a fluorescent ion indicator, whose emission properties change with ion concentration.

Alternatively, the IBL is a molecule whose color or luminescence properties change in the presence of various solvents. For example, the IBL may be a fluorescent molecule such as an ethidium salt whose fluorescence intensity increases in hydrophobic environments. Similarly, the IBL may be a derivative of fluorescein whose color changes between aqueous and nonpolar solvents.

In one embodiment, the DBL may be attached to a bead, i.e. a "decoder bead", that may carry a label such as a fluorophore.

In a preferred embodiment, the IBL-DBL pair comprise substantially complementary single-stranded nucleic acids. In this embodiment, the binding ligands can be referred to as "identifier probes" and "decoder probes". Generally, the identifier and decoder probes range from about 4 basepairs in length to about 1000, with from about 6 to about 100 being preferred, and from about 8 to about 40 being particularly preferred. What is important is that the probes are long enough to be specific, i.e. to distinguish between different IBL-DBL pairs, yet short enough to allow both a) dissociation, if necessary, under suitable experimental conditions, and b) efficient hybridization.

In a preferred embodiment, as is more fully outlined below, the IBLs do not bind to DBLs. Rather, the IBLs are used as identifier moieties ("IMs") that are identified directly, for example through the use of mass spectroscopy.

Alternatively, in a preferred embodiment, the IBL and the capture probe are the same moiety; thus, for example, as outlined herein, particularly when no optical signatures are used, the capture probe can serve as both the identifier and the agent. For example, in the case of nucleic acids, the bead-bound probe (which serves as the capture probe) can also bind decoder probes, to identify the sequence of the probe on the bead. Thus, in this embodiment, the DBLs bind to the capture probes.

In a preferred embodiment, the microspheres may contain an optical signature. That is, as outlined in U.S.S.N.s 08/818,199 and 09/151,877, previous work had each subpopulation of microspheres comprising a unique optical signature or optical tag that is used to identify the unique capture probe of that subpopulation of microspheres; that is, decoding utilizes optical properties of the beads such that a bead comprising the unique optical signature may be distinguished from beads at other locations with different optical signatures. Thus the previous work assigned each capture probe a unique optical

signature such that any microspheres comprising that capture probe are identifiable on the basis of the signature. These optical signatures comprised dyes, usually chromophores or fluorophores, that were entrapped or attached to the beads themselves. Diversity of optical signatures utilized different fluorochromes, different ratios of mixtures of fluorochromes, and different concentrations (intensities) of fluorochromes.

In a preferred embodiment, the present invention does not rely solely on the use of optical properties to decode the arrays. However, as will be appreciated by those in the art, it is possible in some embodiments to utilize optical signatures as an additional coding method, in conjunction with the present system. Thus, for example, as is more fully outlined below, the size of the array may be effectively increased while using a single set of decoding moieties in several ways, one of which is the use of optical signatures one some beads. Thus, for example, using one "set" of decoding molecules, the use of two populations of beads, one with an optical signature and one without, allows the effective doubling of the array size. The use of multiple optical signatures similarly increases the possible size of the array.

In a preferred embodiment, each subpopulation of beads comprises a plurality of different IBLs. By using a plurality of different IBLs to encode each capture probe, the number of possible unique codes is substantially increased. That is, by using one unique IBL per capture probe, the size of the array will be the number of unique IBLs (assuming no "reuse" occurs, as outlined below). However, by using a plurality of different IBLs per bead, n, the size of the array can be increased to 2^n, when the presence or absence of each IBL is used as the indicator. For example, the assignment of 10 IBLs per bead generates a 10 bit binary code, where each bit can be designated as "1" (IBL is present) or "0" (IBL is absent). A 10 bit binary code has 2^{10} possible variants However, as is more fully discussed below, the size of the array may be further increased if another parameter is included such as concentration or intensity; thus for example, if two different concentrations of the IBL are used, then the array size increases as 3^n. Thus, in this embodiment, each individual capture probe in the array is assigned a combination of IBLs, which can be added to the beads prior to the addition of the capture probe, after, or during the synthesis of the capture probe, i.e. simultaneous addition of IBLs and capture probe components.

Alternatively, the combination of different IBLs can be used to elucidate the sequence of the nucleic acid. Thus, for example, using two different IBLs (IBL1 and IBL2), the first position of a nucleic acid can be elucidated: for example, adenosine can be represented by the presence of both IBL1 and IBL2; thymidine can be represented by the presence of IBL1 but not IBL2, cytosine can be represented by the presence of IBL2 but not IBL1, and guanosine can be represented by the absence of both. The second position of the nucleic acid can be done in a similar manner using IBL3 and IBL4; thus, the

presence of IBL1, IBL2, IBL3 and IBL4 gives a sequence of AA; IBL1, IBL2, and IBL3 shows the sequence AT; IBL1, IBL3 and IBL4 gives the sequence TA, etc. The third position utilizes IBL5 and IBL6, etc. In this way, the use of 20 different identifiers can yield a unique code for every possible 10-mer.

In this way, a sort of "bar code" for each sequence can be constructed; the presence or absence of each distinct IBL will allow the identification of each capture probe.

In addition, the use of different concentrations or densities of IBLs allows a "reuse" of sorts. If, for example, the bead comprising a first agent has a 1X concentration of IBL, and a second bead comprising a second agent has a 10X concentration of IBL, using saturating concentrations of the corresponding labelled DBL allows the user to distinguish between the two beads.

Once the microspheres comprising the capture probes are generated, they are added to the substrate to form an array. It should be noted that while most of the methods described herein add the beads to the substrate prior to the assay, the order of making, using and decoding the array can vary. For example, the array can be made, decoded, and then the assay done. Alternatively, the array can be made, used in an assay, and then decoded; this may find particular use when only a few beads need be decoded. Alternatively, the beads can be added to the assay mixture, i.e. the sample containing the target sequences, prior to the addition of the beads to the substrate; after addition and assay, the array may be decoded. This is particularly preferred when the sample comprising the beads is agitated or mixed; this can increase the amount of target sequence bound to the beads per unit time, and thus (in the case of nucleic acid assays) increase the hybridization kinetics. This may find particular use in cases where the concentration of target sequence in the sample is low; generally, for low concentrations, long binding times must be used.

In general, the methods of making the arrays and of decoding the arrays is done to maximize the number of different candidate agents that can be uniquely encoded. The compositions of the invention may be made in a variety of ways. In general, the arrays are made by adding a solution or slurry comprising the beads to a surface containing the sites for attachment of the beads. This may be done in a variety of buffers, including aqueous and organic solvents, and mixtures. The solvent can evaporate, and excess beads removed.

In a preferred embodiment, when non-covalent methods are used to associate the beads to the array, a novel method of loading the beads onto the array is used. This method comprises exposing the array to a solution of particles (including microspheres and cells) and then applying energy, e.g. agitating or vibrating the mixture. This results in an array comprising more tightly associated particles,

as the agitation is done with sufficient energy to cause weakly-associated beads to fall off (or out, in the case of wells). These sites are then available to bind a different bead. In this way, beads that exhibit a high affinity for the sites are selected. Arrays made in this way have two main advantages as compared to a more static loading: first of all, a higher percentage of the sites can be filled easily, and secondly, the arrays thus loaded show a substantial decrease in bead loss during assays. Thus, in a preferred embodiment, these methods are used to generate arrays that have at least about 50% of the sites filled, with at least about 75% being preferred, and at least about 90% being particularly preferred. Similarly, arrays generated in this manner preferably lose less than about 20% of the beads during an assay, with less than about 10% being preferred and less than about 5% being particularly preferred.

In this embodiment, the substrate comprising the surface with the discrete sites is immersed into a solution comprising the particles (beads, cells, etc.). The surface may comprise wells, as is described herein, or other types of sites on a patterned surface such that there is a differential affinity for the sites. This differnetial affinity results in a competitive process, such that particles that will associate more tightly are selected. Preferably, the entire surface to be "loaded" with beads is in fluid contact with the solution. This solution is generally a slurry ranging from about 10,000:1 beads:solution (vol:vol) to 1:1. Generally, the solution can comprise any number of reagents, including aqueous buffers, organic solvents, salts, other reagent components, etc. In addition, the solution preferably comprises an excess of beads; that is, there are more beads than sites on the array. Preferred embodiments utilize two-fold to billion-fold excess of beads.

The immersion can mimic the assay conditions; for example, if the array is to be "dipped" from above into a microtiter plate comprising samples, this configuration can be repeated for the loading, thus minimizing the beads that are likely to fall out due to gravity.

Once the surface has been immersed, the substrate, the solution, or both are subjected to a competitive process, whereby the particles with lower affinity can be disassociated from the substrate and replaced by particles exhibiting a higher affinity to the site. This competitive process is done by the introduction of energy, in the form of heat, sonication, stirring or mixing, vibrating or agitating the solution or substrate, or both.

A preferred embodiment utilizes agitation or vibration. In general, the amount of manipulation of the substrate is minimized to prevent damage to the array; thus, preferred embodiments utilize the agitation of the solution rather than the array, although either will work. As will be appreciated by those in the art, this agitation can take on any number of forms, with a preferred embodiment utilizing microtiter plates comprising bead solutions being agitated using microtiter plate shakers.

The agitation proceeds for a period of time sufficient to load the array to a desired fill. Depending on the size and concentration of the beads and the size of the array, this time may range from about 1 second to days, with from about 1 minute to about 24 hours being preferred.

It should be noted that not all sites of an array may comprise a bead; that is, there may be some sites on the substrate surface which are empty. In addition, there may be some sites that contain more than one bead, although this is not preferred.

In some embodiments, for example when chemical attachment is done, it is possible to attach the beads in a non-random or ordered way. For example, using photoactivatible attachment linkers or photoactivatible adhesives or masks, selected sites on the array may be sequentially rendered suitable for attachment, such that defined populations of beads are laid down.

The arrays of the present invention are constructed such that information about the identity of the capture probe is built into the array, such that the random deposition of the beads in the fiber wells can be "decoded" to allow identification of the capture probe at all positions. This may be done in a variety of ways, and either before, during or after the use of the array to detect target molecules.

Thus, after the array is made, it is "decoded" in order to identify the location of one or more of the capture probes, i.e. each subpopulation of beads, on the substrate surface.

In a preferred embodiment, a selective decoding system is used. In this case, only those microspheres exhibiting a change in the optical signal as a result of the binding of a target sequence are decoded. This is commonly done when the number of "hits", i.e. the number of sites to decode, is generally low. That is, the array is first scanned under experimental conditions in the absence of the target sequences. The sample containing the target sequences is added, and only those locations exhibiting a change in the optical signal are decoded. For example, the beads at either the positive or negative signal locations may be either selectively tagged or released from the array (for example through the use of photocleavable linkers), and subsequently sorted or enriched in a fluorescence-activated cell sorter (FACS). That is, either all the negative beads are released, and then the positive beads are either released or analyzed in situ, or alternatively all the positives are released and analyzed. Alternatively, the labels may comprise halogenated aromatic compounds, and detection of the label is done using for example gas chromatography, chemical tags, isotopic tags mass spectral tags.

As will be appreciated by those in the art, this may also be done in systems where the array is not decoded; i.e. there need not ever be a correlation of bead composition with location. In this

embodiment, the beads are loaded on the array, and the assay is run. The "positives", i.e. those beads displaying a change in the optical signal as is more fully outlined below, are then "marked" to distinguish or separate them from the "negative" beads. This can be done in several ways, preferably using fiber optic arrays. In a preferred embodiment, each bead contains a fluorescent dye. After the assay and the identification of the "positives" or "active beads", light is shown down either only the positive fibers or only the negative fibers, generally in the presence of a light-activated reagent (typically dissolved oxygen). In the former case, all the active beads are photobleached. Thus, upon non-selective release of all the beads with subsequent sorting, for example using a fluorescence activated cell sorter (FACS) machine, the non-fluorescent active beads can be sorted from the fluorescent negative beads. Alternatively, when light is shown down the negative fibers, all the negatives are non-fluorescent and the the postives are fluorescent, and sorting can proceed. The characterization of the attached capture probe may be done directly, for example using mass spectroscopy.

Alternatively, the identification may occur through the use of identifier moieties ("IMs"), which are similar to IBLs but need not necessarily bind to DBLs. That is, rather than elucidate the structure of the capture probe directly, the composition of the IMs may serve as the identifier. Thus, for example, a specific combination of IMs can serve to code the bead, and be used to identify the agent on the bead upon release from the bead followed by subsequent analysis, for example using a gas chromatograph or mass spectroscope.

Alternatively, rather than having each bead contain a fluorescent dye, each bead comprises a non-fluorescent precursor to a fluorescent dye. For example, using photocleavable protecting groups, such as certain ortho-nitrobenzyl groups, on a fluorescent molecule, photoactivation of the fluorochrome can be done. After the assay, light is shown down again either the "positive" or the "negative" fibers, to distinquish these populations. The illuminated precursors are then chemically converted to a fluorescent dye. All the beads are then released from the array, with sorting, to form populations of fluorescent and non-fluorescent beads (either the positives and the negatives or vice versa).

In an alternate preferred embodiment, the sites of attachment of the beads (for example the wells) include a photopolymerizable reagent, or the photopolymerizable agent is added to the assembled array. After the test assay is run, light is shown down again either the "positive" or the "negative" fibers, to distinquish these populations. As a result of the irradiation, either all the positives or all the negatives are polymerized and trapped or bound to the sites, while the other population of beads can be released from the array.

In a preferred embodiment, the location of every capture probe is determined using decoder binding ligands (DBLs). As outlined above, DBLs are binding ligands that will either bind to identifier binding ligands, if present, or to the capture probes themselves, preferably when the capture probe is a nucleic acid or protein.

In a preferred embodiment, as outlined above, the DBL binds to the IBL.

In a preferred embodiment, the capture probes are single-stranded nucleic acids and the DBL is a substantially complementary single-stranded nucleic acid that binds (hybridizes) to the capture probe, termed a decoder probe herein. A decoder probe that is substantially complementary to each candidate probe is made and used to decode the array. In this embodiment, the candidate probes and the decoder probes should be of sufficient length (and the decoding step run under suitable conditions) to allow specificity; i.e. each candidate probe binds to its corresponding decoder probe with sufficient specificity to allow the distinction of each candidate probe.

In a preferred embodiment, the DBLs are either directly or indirectly labeled. In a preferred embodiment, the DBL is directly labeled, that is, the DBL comprises a label. In an alternate embodiment, the DBL is indirectly labeled; that is, a labeling binding ligand (LBL) that will bind to the DBL is used. In this embodiment, the labeling binding ligand-DBL pair can be as described above for IBL-DBL pairs.

Accordingly, the identification of the location of the individual beads (or subpopulations of beads) is done using one or more decoding steps comprising a binding between the labeled DBL and either the IBL or the capture probe (i.e. a hybridization between the candidate probe and the decoder probe when the capture probe is a nucleic acid). After decoding, the DBLs can be removed and the array can be used; however, in some circumstances, for example when the DBL binds to an IBL and not to the capture probe, the removal of the DBL is not required (although it may be desirable in some circumstances). In addition, as outlined herein, decoding may be done either before the array is used to in an assay, during the assay, or after the assay.

In one embodiment, a single decoding step is done. In this embodiment, each DBL is labeled with a unique label, such that the the number of unique tags is equal to or greater than the number of capture probes (although in some cases, "reuse" of the unique labels can be done, as described herein; similarly, minor variants of candidate probes can share the same decoder, if the variants are encoded in another dimension, i.e. in the bead size or label). For each capture probe or IBL, a DBL is made that will specifically bind to it and contains a unique tag, for example one or more fluorochromes. Thus, the identity of each DBL, both its composition (i.e. its sequence when it is a nucleic acid) and its

label, is known. Then, by adding the DBLs to the array containing the capture probes under conditions which allow the formation of complexes (termed hybridization complexes when the components are nucleic acids) between the DBLs and either the capture probes or the IBLs, the location of each DBL can be elucidated. This allows the identification of the location of each capture probe; the random array has been decoded. The DBLs can then be removed, if necessary, and the target sample applied.

In a preferred embodiment, the number of unique labels is less than the number of unique capture probes, and thus a sequential series of decoding steps are used. In this embodiment, decoder probes are divided into **n** sets for decoding. The number of sets corresponds to the number of unique tags. Each decoder probe is labeled in **n** separate reactions with **n** distinct tags. All the decoder probes share the same **n** tags. The decoder probes are pooled so that each pool contains only one of the **n** tag versions of each decoder, and no two decoder probes have the same sequence of tags across all the pools. The number of pools required for this to be true is determined by the number of decoder probes and the **n**. Hybridization of each pool to the array generates a signal at every address. The sequential hybridization of each pool in turn will generate a unique, sequence-specific code for each candidate probe. This identifies the candidate probe at each address in the array. For example, if four tags are used, then 4 X **n** sequential hybridizations can ideally distinguish 4^n sequences, although in some cases more steps may be required. After the hybridization of each pool, the hybrids are denatured and the decoder probes removed, so that the probes are rendered single-stranded for the next hybridization (although it is also possible to hybridize limiting amounts of target so that the available probe is not saturated. Sequential hybridizations can be carried out and analyzed by subtracting pre-existing signal from the previous hybridization).

An example is illustrative. Assuming an array of 16 probe nucleic acids (numbers 1-16), and four unique tags (four different fluors, for example; labels A-D). Decoder probes 1-16 are made that correspond to the probes on the beads. The first step is to label decoder probes 1-4 with tag A, decoder probes 5-8 with tag B, decoder probes 9-12 with tag C, and decoder probes 13-16 with tag D. The probes are mixed and the pool is contacted with the array containing the beads with the attached candidate probes. The location of each tag (and thus each decoder and candidate probe pair) is then determined. The first set of decoder probes are then removed. A second set is added, but this time, decoder probes 1, 5, 9 and 13 are labeled with tag A, decoder probes 2, 6, 10 and 14 are labeled with tag B, decoder probes 3, 7, 11 and 15 are labeled with tag C, and decoder probes 4, 8, 12 and 16 are labeled with tag D. Thus, those beads that contained tag A in both decoding steps contain candidate probe 1; tag A in the first decoding step and tag B in the second decoding step contain candidate probe 2; tag A in the first decoding step and tag C in the second step contain candidate probe 3; etc. In one embodiment, the decoder probes are labeled in situ; that is, they need not be labeled prior to

the decoding reaction. In this embodiment, the incoming decoder probe is shorter than the candidate probe, creating a 5' "overhang" on the decoding probe. The addition of labeled ddNTPs (each labeled with a unique tag) and a polymerase will allow the addition of the tags in a sequence specific manner, thus creating a sequence-specific pattern of signals. Similarly, other modifications can be done, including ligation, etc.

In addition, since the size of the array will be set by the number of unique decoding binding ligands, it is possible to "reuse" a set of unique DBLs to allow for a greater number of test sites. This may be done in several ways; for example, by using some subpopulations that comprise optical signatures. Similarly, the use of a positional coding scheme within an array; different sub-bundles may reuse the set of DBLs. Similarly, one embodiment utilizes bead size as a coding modality, thus allowing the reuse of the set of unique DBLs for each bead size. Alternatively, sequential partial loading of arrays with beads can also allow the reuse of DBLs. Furthermore, "code sharing" can occur as well.

In a preferred embodiment, the DBLs may be reused by having some subpopulations of beads comprise optical signatures. In a preferred embodiment, the optical signature is generally a mixture of reporter dyes, preferably fluoroscent. By varying both the composition of the mixture (i.e. the ratio of one dye to another) and the concentration of the dye (leading to differences in signal intensity), matrices of unique optical signatures may be generated. This may be done by covalently attaching the dyes to the surface of the beads, or alternatively, by entrapping the dye within the bead.

In a preferred embodiment, the encoding can be accomplished in a ratio of at least two dyes, although more encoding dimensions may be added in the size of the beads, for example. In addition, the labels are distinguishable from one another; thus two different labels may comprise different molecules (i.e. two different fluors) or, alternatively, one label at two different concentrations or intensity.

In a preferred embodiment, the dyes are covalently attached to the surface of the beads. This may be done as is generally outlined for the attachment of the capture probes, using functional groups on the surface of the beads. As will be appreciated by those in the art, these attachments are done to minimize the effect on the dye.

In a preferred embodiment, the dyes are non-covalently associated with the beads, generally by entrapping the dyes in the pores of the beads.

Additionally, encoding in the ratios of the two or more dyes, rather than single dye concentrations, is preferred since it provides insensitivity to the intensity of light used to interrogate the reporter dye's signature and detector sensitivity.

In a preferred embodiment, a spatial or positional coding system is done. In this embodiment, there are sub-bundles or subarrays (i.e. portions of the total array) that are utilized. By analogy with the telephone system, each subarray is an "area code", that can have the same tags (i.e. telephone numbers) of other subarrays, that are separated by virtue of the location of the subarray. Thus, for example, the same unique tags can be reused from bundle to bundle. Thus, the use of 50 unique tags in combination with 100 different subarrays can form an array of 5000 different capture probes. In this embodiment, it becomes important to be able to identify one bundle from another; in general, this is done either manually or through the use of marker beads, i.e. beads containing unique tags for each subarray.

In alternative embodiments, additional encoding parameters can be added, such as microsphere size. For example, the use of different size beads may also allow the reuse of sets of DBLs; that is, it is possible to use microspheres of different sizes to expand the encoding dimensions of the microspheres. Optical fiber arrays can be fabricated containing pixels with different fiber diameters or cross-sections; alternatively, two or more fiber optic bundles, each with different cross-sections of the individual fibers, can be added together to form a larger bundle; or, fiber optic bundles with fiber of the same size cross-sections can be used, but just with different sized beads. With different diameters, the largest wells can be filled with the largest microspheres and then moving onto progressively smaller microspheres in the smaller wells until all size wells are then filled. In this manner, the same dye ratio could be used to encode microspheres of different sizes thereby expanding the number of different oligonucleotide sequences or chemical functionalities present in the array. Although outlined for fiber optic substrates, this as well as the other methods outlined herein can be used with other substrates and with other attachment modalities as well.

In a preferred embodiment, the coding and decoding is accomplished by sequential loading of the microspheres into the array. As outlined above for spatial coding, in this embodiment, the optical signatures can be "reused". In this embodiment, the library of microspheres each comprising a different capture probe (or the subpopulations each comprise a different capture probe), is divided into a plurality of sublibraries; for example, depending on the size of the desired array and the number of unique tags, 10 sublibraries each comprising roughly 10% of the total library may be made, with each sublibrary comprising roughly the same unique tags. Then, the first sublibrary is added to the fiber optic bundle comprising the wells, and the location of each capture probe is determined, generally through the use of DBLs. The second sublibrary is then added, and the location of each capture probe is again determined. The signal in this case will comprise the signal from the "first" DBL and the "second" DBL; by comparing the two matrices the location of each bead in each sublibrary can be determined. Similarly, adding the third, fourth, etc. sublibraries sequentially will allow the array to be filled.

In a preferred embodiment, codes can be "shared" in several ways. In a first embodiment, a single code (i.e. IBL/DBL pair) can be assigned to two or more agents if the target sequences different sufficiently in their binding strengths. For example, two nucleic acid probes used in an mRNA quantitation assay can share the same code if the ranges of their hybridization signal intensities do not overlap. This can occur, for example, when one of the target sequences is always present at a much higher concentration than the other. Alternatively, the two target sequences might always be present at a similar concentration, but differ in hybridization efficiency.

Alternatively, a single code can be assigned to multiple agents if the agents are functionally equivalent. For example, if a set of oligonucleotide probes are designed with the common purpose of detecting the presence of a particular gene, then the probes are functionally equivalent, even though they may differ in sequence. Similarly, an array of this type could be used to detect homologs of known genes. In this embodiment, each gene is represented by a heterologous set of probes, hybridizing to different regions of the gene (and therefore differing in sequence). The set of probes share a common code. If a homolog is present, it might hybridize to some but not all of the probes. The level of homology might be indicated by the fraction of probes hybridizing, as well as the average hybridization intensity. Similarly, multiple antibodies to the same protein could all share the same code.

In a preferred embodiment, decoding of self-assembled random arrays is done on the bases of pH titration. In this embodiment, in addition to capture probes, the beads comprise optical signatures, wherein the optical signatures are generated by the use of pH-responsive dyes (sometimes referred to herein as "ph dyes") such as fluorophores. This embodiment is similar to that outlined in PCT US98/05025 and U.S.S.N. 09/151,877, both of which are expressly incorporated by reference, except that the dyes used in the present ivention exhibits changes in fluorescence intensity (or other properties) when the solution pH is adjusted from below the pKa to above the pKa (or vice versa). In a preferred embodiment, a set of pH dyes are used, each with a different pKa, preferably separated by at least 0.5 pH units. Preferred embodiments utilize a pH dye set of pKa's of 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0, 10.5, 11, and 11.5. Each bead can contain any subset of the pH dyes, and in this way a unique code for the capture probe is generated. Thus, the decoding of an array is achieved by titrating the array from pH 1 to pH 13, and measuring the fluorescence signal from each bead as a function of solution pH.

Thus, the present invention provides array compositions comprising a substrate with a surface comprising discrete sites. A population of microspheres is distributed on the sites, and the population comprises at least a first and a second subpopulation. Each subpopulation comprises a capture probe, and, in addition, at least one optical dye with a given pKa. The pKas of the different optical dyes are different.

In a preferred embodiment, "random" decoding probes can be made. By sequential hybridizations or the use of multiple labels, as is outlined above, a unique hybridization pattern can be generated for each sensor element. This allows all the beads representing a given clone to be identified as belonging to the same group. In general, this is done by using random or partially degenerate decoding probes, that bind in a sequence-dependent but not highly sequence-specific manner. The process can be repeated a number of times, each time using a different labeling entity, to generate a different pattern of signals based on quasi-specific interactions. In this way, a unique optical signature is eventually built up for each sensor element. By applying pattern recognition or clustering algorithms to the optical signatures, the beads can be grouped into sets that share the same signature (i.e. carry the same probes).

In order to identify the actual sequence of the clone itself, additional procedures are required; for example, direct sequencing can be done, or an ordered array containing the clones, such as a spotted cDNA array, to generate a "key" that links a hybridization pattern to a specific clone.

Alternatively, clone arrays can be decoded using binary decoding with vector tags. For example, partially randomized oligos are cloned into a nucleic acid vector (e.g. plasmid, phage, etc.). Each oligonucleotide sequence consists of a subset of a limited set of sequences. For example, if the limites set comprises 10 sequences, each oligonucleotide may have some subset (or all of the 10) sequences. Thus each of the 10 sequences can be present or absent in the oligonucleotide. Therefore, there are 2^{10} or 1,024 possible combinations. The sequences may overlap, and minor variants can also be represented (e.g. A, C, T and G substitutions) to increase the number of possible combinations. A nucleic acid library is cloned into a vector containing the random code sequences. Alternatively, other methods such as PCR can be used to add the tags. In this way it is possible to use a small number of oligo decoding probes to decode an array of clones.

In a preferred embodiment, pyrosequencing techniques as described above are used to decode the array. That is, pyrosequencing is used to identify or sequence the DBL on each bead of the array. Accordingly, the array is decoded.

An advantage of using array formats such as have been described is that minimal reagents are required for the different sequencing reactions described. For example, methods based on the addition (sequencing by synthesis) of nucleotides requires multiple changes of reagent coupled with intervening reading steps. When sequencing templates are immobilized on beads and are associated with a substrate such as a fiber optic bundle, the fiber optic bundle can be contacted with different reagents, such as a well containing the nucleotide "A". Upon imaging the incorporation of the

nucleotide with the imaging system at the distal end of the fiber bundle, the fiber optic bundle containing the immobilized sequencing template is removed from the first reagent, optionally washed in a second well containing a wash solution, and placed in a third well containing a different substance, such as the nucleotide "T". This cycle can be repeated as necessary to generate a sequence of the sequencing template. As such, many individual reactions are performed in parallel on each array.

In a preferred embodiment, multiple arrays are processed in parallel. For example, a distinct fiber optic array comprising different sequencing templates can be in each of the four nucleotides (A, T, G and C) at the same time, and analyzed simultaneously. Thus, upon imaging the result of one sequencing reaction with a particular nucleotide, each of the fiber optic bundles is moved to one of the remaining wells containing an as yet unexamined nucleotide.

Advantages of the system include the ability to include washing or other processing steps that may be carried out between sequencing reaction. This is accomplished by dipping the fiber into a well containing the appropriate reagents. Again, many individual reactions are carried out in parallel on each array. Multiple arrays can be processed in parallel. By bringing the array to the reagents in a well, fluid handling is made simple and efficient. By using an optical imaging fiber, imaging can be carried out conveniently in real time, which greatly facilitates time-resolved sequencing.

The system also lends itself to automation. For example, in one format four reaction wells are arranged in a circular format, for the reactions with "A", " C", "G" and "T" nucleotides. Four fibers are arranged so that they can dip into the four wells simultaneously. After carrying out a cycle of sequencing and reading, a 90 degree rotation is carried out, so that each fiber dips into the next reaction well. In this way, a continuous series of sequencing reactions can be carried out on the multiple fibers in parallel. Intervening wells can also be used for other processing steps if required. For example, if it is necessary to use a reagent that must be replenished at each step, variations on this design can be used. For example, a mechanism can be included for recharging the wells. Alternatively, the fibers can be dipped into a continuous stream of reagent.

Another preferred embodiment couples sequencing by synthesis with oligonucleotide synthesis. A disadvantage of sequencing by synthesis is that relatively few positions can be analyzed (e.g. 20 -50 nucleotides with reasonable accuracy with pyrosequencing). Thus, an additional stepwise procedure is included in the sequencing format. That is, after a template nucleic acid has been sequenced using sequencing by synthesis reactions, the new sequence information is automatically used to design a new sequencing primer. The primer sequence can be transferred to an oligonucleotide synthesizer, for example a small-scale oligonucleotide synthesizer, which generates a new primer that is used for a second round of sequencing by synthesis. In this way multiple overlapping rounds of sequencing by

synthesis are used to generate long sequences from template nucleic acids. Ideally, the processing steps (sequencing and oligonucleotide synthesis) and information processing are fully integrated, so that long sequence reads are obtained automatically. Because of the small dimensions of the self-assembled arrays on optical fibers, this is possible using microfluidic fluid transport and processing.

As previously described a nucleic acid molecule attached to the bead acts as the bead identifier (it may also provide the sensor function of the bead). By using time-resolved sequencing to read out the sequence of the nucleic acid, the bead is decoded. Thus, using the methods described above, time-resolved sequencing can be used to decode a self-assembled array in a highly parallel way.

Detection of the sequencing reactions of the invention, including the direct detection of sequencing products and indirect detection utilizing label probes (i.e. sandwich assays), is done by detecting assay complexes comprising labels.

In a preferred embodiment, several levels of redundancy are built into the arrays of the invention. Building redundancy into an array gives several significant advantages, including the ability to make quantitative estimates of confidence about the data and signficant increases in sensitivity. Thus, preferred embodiments utilize array redundancy. As will be appreciated by those in the art, there are at least two types of redundancy that can be built into an array: the use of multiple identical sensor elements (termed herein "sensor redundancy"), and the use of multiple sensor elements directed to the same target analyte, but comprising different chemical functionalities (termed herein "target redundancy"). For example, for the detection of nucleic acids, sensor redundancy utilizes of a plurality of sensor elements such as beads comprising identical binding ligands such as probes. Target redundancy utilizes sensor elements with different probes to the same target: one probe may span the first 25 bases of the target, a second probe may span the second 25 bases of the target, etc. By building in either or both of these types of redundancy into an array, significant benefits are obtained. For example, a variety of statistical mathematical analyses may be done.

In addition, while this is generally described herein for bead arrays, as will be appreciated by those in the art, this techniques can be used for any type of arrays designed to detect target analytes. Furthermore, while these techniques are generally described for nucleic acid systems, these techniques are useful in the detection of other binding ligand/target analyte systems as well.

In a preferred embodiment, sensor redundancy is used. In this embodiment, a plurality of sensor elements, e.g. beads, comprising identical bioactive agents are used. That is, each subpopulation comprises a plurality of beads comprising identical bioactive agents (e.g. binding ligands). By using a

number of identical sensor elements for a given array, the optical signal from each sensor element can be combined and any number of statistical analyses run, as outlined below. This can be done for a variety of reasons. For example, in time varying measurements, redundancy can significantly reduce the noise in the system. For non-time based measurements, redundancy can significantly increase the confidence of the data.

In a preferred embodiment, a plurality of identical sensor elements are used. As will be appreciated by those in the art, the number of identical sensor elements will vary with the application and use of the sensor array. In general, anywhere from 2 to thousands may be used, with from 2 to 100 being preferred, 2 to 50 being particularly preferred and from 5 to 20 being especially preferred. In general, preliminary results indicate that roughly 10 beads gives a sufficient advantage, although for some applications, more identical sensor elements can be used.

Once obtained, the optical response signals from a plurality of sensor beads within each bead subpopulation can be manipulated and analyzed in a wide variety of ways, including baseline adjustment, averaging, standard deviation analysis, distribution and cluster analysis, confidence interval analysis, mean testing, etc.

Once the baseline has been adjusted, a number of possible statistical analyses may be run to generate known statistical parameters. Analyses based on redundancy are known and generally described in texts such as Freund and Walpole, Mathematical Statistics, Prentice Hall, Inc. New Jersey, 1980, hereby incorporated by reference in its entirety.

In a preferred embodiment, signal summing is done by simply adding the intensity values of all responses at each time point, generating a new temporal response comprised of the sum of all bead responses. These values can be baseline-adjusted or raw. As for all the analyses described herein, signal summing can be performed in real time or during post-data acquisition data reduction and analysis.

In a preferred embodiment, cummulative response data is generated by simply adding all data points in successive time intervals. This final column, comprised of the sum of all data points at a particular time interval, may then be compared or plotted with the individual bead responses to determine the extent of signal enhancement or improved signal-to-noise ratios.

In a preferred embodiment, the mean of the subpopulation (i.e. the plurality of identical beads) is determined, using the well known Equation 1:

Equation 1

$$\mu = \sum \frac{x_i}{n}$$

In some embodiments, the subpopulation may be redefined to exclude some beads if necessary (for example for obvious outliers, as discussed below).

In a preferred embodiment, the standard deviation of the subpopulation can be determined, generally using Equation 2 (for the entire subpopulation) and Equation 3 (for less than the entire subpopulation):

Equation 2

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

Equation 3

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

As for the mean, the subpopulation may be redefined to exclude some beads if necessary (for example for obvious outliers, as discussed below).

In a preferred embodiment, statistical analyses are done to evaluate whether a particular data point has statistical validty within a subpopulation by using techniques including, but not limited to, t distribution and cluster analysis. This may be done to statistically discard outliers that may otherwise skew the result and increase the signal-to-noise ratio of any particular experiment. This may be done using Equation 4:

Equation 4

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

In a preferred embodiment, the quality of the data is evaluated using confidence intervals, as is known

in the art. Confidence intervals can be used to facilitate more comprehensive data processing to measure the statistical validity of a result.

In a preferred embodiment, statistical parameters of a subpopulation of beads are used to do hypothesis testing. One application is tests concerning means, also called mean testing. In this application, statistical evaluation is done to determine whether two subpopulations are different. For example, one sample could be compared with another sample for each subpopulation within an array to determine if the variation is statistically significant.

In addition, mean testing can also be used to differentiate two different assays that share the same code. If the two assays give results that are statistically distinct from each other, then the subpopulations that share a common code can be distinguished from each other on the basis of the assay and the mean test, shown below in Equation 5:

Equation 5

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Furthermore, analyzing the distribution of individual members of a subpopulation of sensor elements may be done. For example, a subpopulation distribution can be evaluated to determine whether the distribution is binomial, Poisson, hypergeometric, etc.

In addition to the sensor redundancy, a preferred embodiment utilizes a plurality of sensor elements that are directed to a single target analyte but yet are not identical. For example, a single target nucleic acid analyte may have two or more sensor elements each comprising a different probe. This adds a level of confidence as non-specific binding interactions can be statistically minimized. When nucleic acid target analytes are to be evaluated, the redundant nucleic acid probes may be overlapping, adjacent, or spatially separated. However, it is preferred that two probes do not compete for a single binding site, so adjacent or separated probes are preferred. Similarly, when proteinaceous target analytes are to be evaluated, preferred embodiments utilize bioactive agent binding agents that bind to different parts of the target. For example, when antibodies (or antibody fragments) are used as bioactive agents for the binding of target proteins, preferred embodiments utilize antibodies to different epitopes.

In this embodiment, a plurality of different sensor elements may be used, with from about 2 to about

20 being preferred, and from about 2 to about 10 being especially preferred, and from 2 to about 5 being particularly preferred, including 2, 3, 4 or 5. However, as above, more may also be used, depending on the application.

As above, any number of statistical analyses may be run on the data from target redundant sensors.

One benefit of the sensor element summing (referred to herein as "bead summing" when beads are used), is the increase in sensitivity that can occur.

In addition, the present invention provides kits comprising the compositions of the invention. In a preferred embodiment, the kit for nucleic acid sequencing comprises an array composition. Preferred embodiments utilize a substrate with a surface comprising discrete sites and a population of microspheres distributed on the sites. However, in some embodiments, the array may not be formulated; that is, the beads may not yet be associated on the surface, which may be done by the end-user. The beads preferably comprise capture probes. The kit additionally comprises at least a first enzyme comprising an extension enzyme, and dNTPs. These may be labelled or unlabelled, derivatized (i.e. protected) or not, depending on the sequencing method and configuration of the system, as is outlined herein.

In some embodiments, the kits may also comprise decoding probes, as described herein.

In a preferred embodiment, the kits contain additional components directed to the sequencing method of choice. For example, preferred embodiments utilize the enzymes and reactants required for pyrosequencing, including, but not limited to, a second enyzme for the conversion of PPi into ATP, a third enzyme for the detection of ATP, and the associated reagents required for the enzymes.

In a preferred embodiment, the kits comprise the components for reversible chain termination sequencing. In this embodiment, the dNTPs comprise a reversible protecting group as outlined herein.

Once made, the methods and compositions of the invention find use in a number of applications. In a preferred embodiment, the sequencing methods find use in the decoding of randomly assembled arrays, particularly bead arrays, as described herein.

In a preferred embodiment, the methods and compositions of the invention find use in sequencing target nucleic acids, which may be done for a wide variety of purposes, as will be appreciated by those in the art. For example, novel genes and regulatory sequences, all or part of any number of genomes

can be sequenced or resequenced using the present invention.

All references cited herein are incorporated by reference in their entirety.